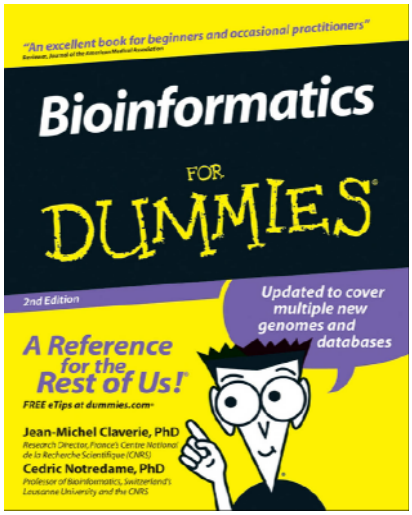


بیوانفورماتیک به زبان ساده

داور مجله انجمن پزشکی آمریکا: "یک کتاب عالی برای مبتدیان و نیمه حرفه ای ها"

چاپ دوم (به روز شده برای ژنومها و داده-پایگاههای جدید)

تالیف: ژان میشل کلاوری (PhD) و سدریک نوتردام (PhD)

ترجمه: حسن کربکندی(PhD)، سجاد عباسی (Pharm D)،

سیاوش ایروانی (Pharm D)، محمود حسینی

پیشگفتار

بیوانفورماتیک یکی از رشته های جدید، مطرح و به سرعت در حال رشد است. این رشته بعلت پرداختن به بعضی مطالب ریاضی و آماری ممکن است مشکل بنظر برسد. بهمین خاطر کمپانی Wiley، یکی از بزرگترین و معروفترین ناشرین علمی جهان و ناشر سلسله کتابهای پرفروش"® For Dummies" که معمولاً کتابهای مربوط به علوم کمی پیچیده مانند کامپیوتر و برنامه نویسی را به زبان ساده به چاپ می رساند، اقدام به انتشار کتابی در این زمینه نیز نموده است.

این کتاب بر خلاف بیشتر کتابهایی که در زمینه بیوانفورماتیک نوشته شده است، به زبانی ساده و بصورت خود آموز نگارش یافته است. مطالب کتاب نیز کاملاً عملی و ۱۰۰٪ کاربردی است و به اصطلاح بصورت Hands on Keyboard قابل استفاده می باشد. از بیان مطالب غامض و غیر کاربردی نیز در این کتاب اجتناب شده است.

اینجانب برای تدریس درس بیوانفورماتیک در کلاسها و کارگاهها، مدت زیادی بدنبال کتاب مناسبی می گشتم و بالاخره این کتاب را ساده، خودآموز و عملی یافتم و مدت ۴ سال است آن را بعنوان کتاب درسی این رشته تدریس می نمایم. این کتاب می تواند برای مقاطع مختلف کارشناسی، کارشناسی ارشد، و دکترای رشته های مختلف بیوتکنولوژی، ژنتیک، مهندسی ژنتیک، ژنتیک انسانی، پزشکی، پزشکی مولکولی، داروسازی، بیوشیمی، میکروبیولوژی، ایمونولوژی و انگل شناسی مورد استفاده قرار گیرد.

در ترجمه سعی شده است لحن ساده، طنز آلود، محاوره ای و دوستانه کتاب حفظ گردد. اما ممکن است مانند سایر ترجمه ها خالی از اشتباه نباشد. لذا ارسال اصلاحات احتمالی شما دانشجویان و اساتید ارجمند در چاپهای بعدی می تواند بسیار راهگشا باشد. از آنجا که اینترنت و سایتهای مربوط به بیوانفورماتیک هر روز در حال تغییر هستند ممکن است در شکل و عمل پاره ای از آنها تغییراتی حاصل گردد و این موضوع اجتناب ناپذیر است. اما تغییرات احتمالی در حدی خواهند بود که کاربر می تواند با تلاش اندکی موفق شود مسیر استفاده خود را در لابلای آنها بیابد.

دکتر حسن کربکندی
PhD بیوتکنولوژی از انگلستان و
استادیار گروه ژنتیک و بیولوژی مولکولی دانشکده پزشکی
دانشگاه علوم پزشکی اصفهان

فهرست مطالب

عنوان | صفحه

مقدمه ۱
فرضیات و قوانین ۲
شمای کلی از مطالب کتاب ۲
بخش اول: شروع کار با بیوانفورماتیک ۲
بخش دوم: راهنمایی های ضروری در خصوص بیوانفورماتیک ۲
بخش سوم: در آنالیز توالی های مهارت کافی پیدا نمایید! ۳
بخش چهارم: یک متخصص شوید!! تکنیک های بیوانفورماتیک پیشرفته ۳
بخش پنجم: قسمت دهگانه ۳
علامت های استفاده شده در این کتاب ۴
هدف نهایی این کتاب ۴
بخش اول: شروع کار با بیوانفورماتیک ۶
فصل اول: آنچه بیوانفورماتیک می تواند برای ما انجام دهد ۸
بیوانفورماتیک چیست؟ ۸
آنالیز توالی های پروتئینی ۹
یک تاریخچه مختصر از آنالیز توالی ۱۱
خواندن توالی های پروتئینی از N تا C ۱۲
کار کردن با ساختار های سه بعدی پروتئین ها ۱۳
سرفصل های مرتبط با بیوانفورماتیک در این کتاب ۱۴
آنالیز توالی های DNA ۱۵
خواندن صحیح توالی های DNA ۱۵
دو شکل خواندن توالی DNA ۱۶
کد آیوپاک نوکلئوتید های DNA ۱۶
توالی های DNA که از دو طرف یکسان خوانده می شوند ۱۷

آنالیز توالی های RNA ۱۸
ساختار RNA: بازی کردن با رشته های چسبنده ۱۹
اطلاعات بیشتر در مورد نامگذاری اسیدهای نوکلئیک ۲۰
نواحی کد کننده پروتئین: شبیه سازی کار با توالی های آمینواسیدی ۲۰
تبدیل DNA به پروتئین: کد ژنتیکی ۲۱
در مورد کد شدن توالی DNA به پروتئین بیشتر بدانیم ۲۲
سرفصل های مرتبط با بیوانفورماتیک DNA و RNA در این کتاب ۲۳
کار کردن با ژنوم کامل ۲۳
ژنومیک: بدست آوردن همزمان همه ژن ها یا ژنوم کامل ۲۴
سرفصل های مرتبط با بیوانفورماتیک ژنوم ۲۵
فصل دوم: چگونه از بیوانفورماتیک استفاده کنیم؟ ۲۶
چگونه می توان به سرعت یک متخصص PubMed/Medline شد؟ ۲۷
گرفتن اطلاعات در مورد یک پروتئین با استفاده از نام آن ۲۷
جستجو کردن در PubMed ۲۷
چگونه چند نتیجه جستجوی دلخواه را ذخیره کنیم؟ ۲۹
جستجو در PubMed با استفاده از نام مؤلف ۲۹
جستجو در PubMed با استفاده از "field" ها ۳۲
استفاده درست از "field" ها ۳۳
استفاده از "field" ها برای یافتن متخصصین یک زمینه خاص ۳۴
اعمال محدودیت در جستجو در PubMed ۳۴
چند نکته دیگر در مورد PubMed ۳۷
بازیابی توالی های پروتئینی ۳۸
ExPASy: یک سایت اینترنتی مناسب برای بدست آوردن اطلاعات در مورد پروتئین ۳۹
راه های پیشرفته تر برای بازیابی توالی پروتئین ها ۴۲
فرمت FASTA و RAW ۴۳
بدست آوردن توالی های پروتئینی مرتبط با هم ۴۴
بازیابی توالی های DNA ۴۷
همه بخش های مولکول DNA به پروتئین کد نمی شوند ۴۷
چگونه هنگام دانلود فایل های توالی ها، از کاراکترهای مزاحم در امان باشیم؟ ۴۸

دستیابی به توالی DNA از روی توالی پروتئین ۴۹
یافتن توالی DNA مربوط به یک پروتئین خاص ۴۹
مقایسه یک توالی پروتئین با توالی های دیگر با استفاده از BLAST ۵۳
تطبیق چند توالی پروتئین با استفاده از برنامه ClustalW ۵۷
بخش دوم: راهنمایی های ضروری در بیوانفورماتیک ۶۱
فصل سوم: استفاده از بانک های اطلاعاتی توالی های نوکلئوتیدی ۶۳
مطالعه در زمینه ژن ها و ژنوم ۶۴
پروکاریوت ها: موجودات کوچک، ژن های ساده ۶۴
یوکاریوت ها: موجودات بزرگ تر، ژن های پیچیده تر ۶۶
نحوه استفاده از GenBank ۶۷
درک صحیح از ورودی های GenBank پروکاریوتی ۶۷
خواندن سربرگ یک ورودی GenBank پروکاریوتی ۶۸
استفاده از قسمت Sequence یک ورودی پروکاریوتی ۷۰
درک مفهوم ورودی یک mRNA یوکاریوتی در GenBank ۷۱
درک مفهوم ورودی یک ژنوم یوکاریوتی در GenBank ۷۳
کار با ورودی های وابسته به هم در GenBank ۷۷
دسترسی به ورودی های بدون شماره های دسترسی در GenBank ۷۸
استفاده از بانک های اطلاعاتی ژن محور ۸۰
کار با بانک های اطلاعاتی ژنوم های کامل ۸۲
کار با ژنوم های کامل ویروسی ۸۲
کار با ژنوم های کامل باکتریایی ۸۵
ژنوم های باکتریایی دیگر در TIGR ۸۸
میکروب های موجود در محیط زیست در DoE ۸۸
کشف ژنوم انسان ۹۰
کسب اطلاعات بیشتر پیرامون پروژه Ensembl ۹۱
شروع کار با سایت Ensembl ۹۱
دستیابی به کارت شناسایی ژن DUT انسانی ۹۵
دستیابی به ژن های بیماری زای حاوی SNP های کد کننده از طریق BioMart ۹۶
فصل چهارم: چگونگی استفاده از بانک های اطلاعاتی توالی و پروتئین ۹۹

از ORF های ترجمه شده تا پروتئین کامل ۱۰۱
ORF ها: همواره آن چیزی که می بینی همان چیزی نیست که به دست می آوری! ۱۰۱
مقصد نهایی هر پروتئین ۱۰۳
گوناگونی و تنوع ترکیبی عملکرد و نحوه تا خوردن ۱۰۳
مطالعه ی که ورودی Swiss-Prot ۱۰۴
کشف رموز ورودی EGFR در Swiss-Prot ۱۰۴
اطلاعات کلی پیرامون ورودی مورد نظر ۱۰۵
نام و منشا یک پروتئین ۱۰۶
منابع ۱۰۷
بخش Comments ۱۰۸
بخش Cross-References ۱۰۹
کلید واژه ها ۱۱۱
بخش Features ۱۱۲
در نهایت، خود توالی ۱۱۵
اطلاعات بیشتر پیرامون پروتئین انتخابی شما ۱۱۵
یادگیری بیشتر پیرامون آمینواسید های تغییر یافته ۱۱۶
معرفی چند سایت تخصصی و پیشرفته بیوشیمی ۱۱۷
اطلاعات بیشتر پیرامون مسیر های بیوشیمیایی ۱۱۷
یافتن اطلاعات بیشتر در مورد ساختار های پروتئین ۱۱۸
یافتن اطلاعات بیشتر پیرامون خانواده های اصلی پروتئین ها ۱۱۹
فصل پنجم: کار با توالی منفرد DNA ۱۲۱
جلوی اشتباه را قبل از آنکه دیر شود بگیرید! ۱۲۲
جدا سازی و حذف توالی های حامل ۱۲۲
مواردی که نباید توالی هایتان را دور بیندازید ۱۲۴
پردازش و بررسی یک Restriction Map ۱۲۵
طراحی پرایمر های PCR ۱۲۷
آنالیز ساختار و ترکیبات DNA ۱۲۹
تخمین محتوای G+C در توالی ۱۳۰
شمارش حروف در توالی DNA ۱۳۰

شمارش کلمات بلند تر در توالی های DNA ۱۳۱
کار کردن با دیگر آنالیز های ترکیبی DNA ۱۳۳
یافتن تکرار های داخلی در یک توالی ۱۳۳
یافتن تکرار ها یک کار ظریف است! ۱۳۳
استفاده از نمودار نقطه ای برای شناسایی تکرار های معکوس ۱۳۵
شناسایی تکرار های مختص ژنوم در یک توالی ۱۳۶
شناسایی مناطق کد کننده ی پروتئین ۱۳۶
ORFing توالی DNA ۱۳۷
آنالیز DNA از طریق GeneMark ۱۳۹
یافتن اگزون های داخلی در توالی های ژنومی مهره داران ۱۴۰
تجزیه کامل ژن برای ژنوم های یوکاریوتی ۱۴۲
آنالیز یک توالی با برنامه GenomeScan ۱۴۲
تجمع قطعات یک توالی ۱۴۳
مدیریت پروژه های توالی یابی بلند توسط نرم افزارهای معروف ۱۴۴
تجمع توالی های مورد نظر از طریق CAP3 ..... ۱۴۵
فراتر از این فصل ۱۴۷
فصل ششم: کار با توالی پروتئینی ۱۴۹
انجام عملیات بیوشیمیائی از طریق کامپیوتر ۱۵۰
شناسائی و پیش بینی خصوصیات فیزیکوشیمیائی یک پروتئین ۱۵۱
تفسیر نتایج حاصل از Prot Param ۱۵۳
وزن مولکولی ۱۵۴
ضرایب خاموش سازی (Extinction coefficients) ۱۵۴
ناپایداری ۱۵۴
نیمه عمر ۱۵۴
تجزیه یک پروتئین در کامپیوتر ۱۵۴
انجام آنالیز ساختار اولیه (primary structure) ۱۵۵
جستجوی قطعات بین غشائی ۱۵۷
اجرای برنامه Protscale ۱۵۷
تفسیر نتایج حاصل از Protscale ۱۵۸
اجرای برنامه TMHMM ۱۵۹

تفسیر نتایج بدست آمده از TMHMM ۱۶۰
جستجوی مناطق فنر مانند (Coiled- coil) ۱۶۱
پیش بینی تغییرات ایجاد شده بر روی یک پروتئین ۱۶۲
جستجوی الگوهای PROSITE ۱۶۳
تفسیر نتایج حاصل از Scan Prosite ۱۶۵
درک صحیح از خروجی Scan Prosite ۱۶۵
مراقب الگوهای(pattern) کوتاه باشید ۱۶۶
استفاده از اطلاعات گونه ها ۱۶۶
سیگنالهای ضعیف با هم ایجاد سیگنالهای قوی تر می کنند ۱۶۷
حذف الگوهای ضعیف ۱۶۷
همه ی مطالب را در Prosite نمی توان یافت ۱۶۸
یافتن دامین های معروف و شناخته شده در پروتئین ۱۶۸
انتخاب مجموعه ی صحیحی از دامین ها ۱۶۸
یافتن دامین ها با استفاده از برنامه Inter ProScan ۱۷۰
تفسیر نتایج InterProScan ۱۷۲
درک صحیح از جزئیات مربوط به خروجی Inter ProScan ۱۷۲
درک صحیح از ناسازگاری ها و ناهماهنگی های احتمالی در خروجی بدست آمده از InterProScan ۱۷۳
یافتن دامین ها از طریق سرور CD ۱۷۴
تفسیر و درک صحیح نتایج حاصل از سرور CD ۱۷۵
یافتن دامین ها با استفاده Motif Scan ۱۷۷
تفسیر و درک صحیح نتایج حاصل از برنامه Motif Scan ۱۷۸
استنتاج صحیح از خروجی Motif Scan ۱۸۰
کشف دامین های جدید در پروتئین ۱۸۰
روش های دیگر برای آنالیز پروتئین از طریق اینترنت ۱۸۱
بخش سوم: در آنالیز توالی ها مهارت کافی پیدا نمائید! ۱۸۲
فصل هفتم: جستجوی توالی های مشابه در بانک های اطلاعاتی ۱۸۴
چرا تشابهات مهمند؟ ۱۸۵
محبوب ترین منبع به دست آوردن داده: BLAST ۱۸۶
BLAST توالی های پروتیینی ۱۸۶
کار کردن با NCBI Blastp ۱۸۷

کار با موتور جستجوگر EMBnet blastp ۱۹۲
چگونه نتایج حاصل از جستجو در یک موتور BLAST را بفهمیم؟ ۱۹۴
نمایی کلی از خروجی BLAST ۱۹۵
نمودارهای ترسیمی ۱۹۵
لیست هدف ۱۹۶
هم ترازی ها (alignments) ۱۹۷
پارامترها ۱۹۹
BLAST توالی های DNA ۲۰۰
کدام نوع BLAST برای توالی DNA ی من بهتر است؟ ۲۰۰
بهترین انتخاب ها برای DNA BLAST چه هستند؟ ۲۰۱
استفاده از روش BLAST برای انجام کارها ۲۰۱
تنظیمات BLAST: انتخاب پارامترهای مناسب ۲۰۳
چگونه به سیستم BLAST بگوییم که کدام مناطق از توالی ها را از محاسبات خود خارج کند؟ ۲۰۴
نقابدار کردن توالی های پروتیینی ۲۰۴
نقابدار کردن توالی های DNA ۲۰۵
تغییر در پارامترهای مرتبط با Alignment ۲۰۶
تنظیم داده های خروجی BLAST ۲۰۶
انتخاب بانک اطلاعاتی متناسب با نیازمان ۲۰۶
محدود کردن نتایج با استفاده از Enterez query ۲۰۷
Expect ۲۰۷
به وسیله ی BLAST-PSI جستجوی خود را تکرار کنیم ۲۰۸
BLAST-PSI توالی های پروتیینی ۲۰۸
چگونه توالی های صحیح را انتخاب کنیم؟ ۲۱۰
چگونه مناطق موجود در پروتیین خود را با استفاده از BLAST و BLAST-PSI شناسایی و جدا کنیم؟ ۲۱۱
جستجوی توالی های مشابه با سرویسهای مجانی در اینترنت ۲۱۱
فصل هشتم: مقایسه کردن دو توالی ۲۱۴
مطمئن شوید که توالی ها و روشهای صحیح را در دست دارید ۲۱۵
انتخاب توالی های صحیح ۲۱۵
انتخاب متد صحیح ۲۱۶
ایجاد یک dotplot ۲۱۶

انتخاب نوع درست Dot Plot ۲۱۹
استفاده از Dotlet در اینترنت ۲۱۹
دانلود کردن Dotlet ۲۲۰
وارد کردن توالی ها در Dotlet ۲۲۱
Dot let تان را درست تنظیم کنید ۲۲۳
تفسیر نتایج ۲۲۶
انجام آنالیزهای بیولوژیک با dotplot ۲۲۷
تشخیص توالی های پشت سر هم ۲۲۷
پیدا کردن نواحی با پیچیدگی کمتر (Low-complexity) ۲۲۹
آنالیز اسیدهای نوکلئیک یا dotlet ۲۳۰
ساخت Local Alignment در اینترنت ۲۳۰
انتخاب global alignment مناسب ۲۳۱
استفاده از Lalign برای پیدا کردن ده global alignment برتر ۲۳۲
تفسیر نتایج حاصل از Lalign ۲۳۴
ساخت Global Alignment در اینترنت ۲۳۶
استفاده از Lalign برای پیدا کردن ده global alignment برتر ۲۳۶
هم ترازی پروتئین ها و DNA ۲۳۷
مقایسه دو به دو توالی به طور رایگان در اینترنت ۲۳۷
فصل نهم: چگونه یک Multiple sequence alignment بسازیم؟ ۲۳۹
چگونه موقعیت هایی که Multiple sequence alignment کمک کننده نخواهد بود را تشخیص دهیم؟ ۲۴۰
چگونه می توان با Multiple sequence alignment به سیر تحقیقات خود کمک کرد؟ ۲۴۱
چگونه مناسب ترین توالی ها را انتخاب کنیم؟ ۲۴۴
انواع توالی هایی که شما در پی آن خواهید بود کدامند؟ ۲۴۴
بهتر است که از پروتیین استفاده کنیم و یا از DNA؟ ۲۴۵
چگونه تعداد مناسب توالی را انتخاب کنیم؟ ۲۴۵
ترکیب مناسب شباهت و اطلاعات جدید ۲۴۶
چگونه توالیهایمان را باروشی صحیح نام گذاری کنیم؟ ۲۴۸
چگونه توالی های مورد نیازمان را با استفاده از سرورهای آنلاین BLAST جمع آوری کنیم؟ ۲۴۸

چگونه توالی های مورد نیاز خود را با استفاده از سرور ExPASy انتخاب کنیم؟ ۲۴۹
چگونه توالی هایی معین را از Swiss-Prot بیابیم؟ ۲۵۱
چگونه روشی مناسب برای Multiple sequence alignment انتخاب کنیم؟ ۲۵۳
چگونه از ClastalW استفاده کنیم؟ ۲۵۴
چگونه از EBI ClatalW استفاده کنیم؟ ۲۵۴
چگونه می توانیم پارامترهای ClustalW را تغییر دهیم؟ ۲۵۶
چگونه با استفاده از T-coffee توالی ها و ساختارها را هم تراز کنیم؟ ۲۵۶
چگونه با استفاده از T-coffee یک multiple sequence alignment بسازیم؟ ۲۵۷
چگونه با استفاده از EXPRESSO توالی ها و ساختارها را با یکدیگر ترکیب کنیم؟ ۲۵۹
چگونه کیفیت alignment خود را با استفاده از CORE ارزیابی کنیم؟ ۲۵۹
چگونه با استفاده از MUSCLE پایگاه های داده ای بزرگ را کوچک کنیم؟ ۲۵۹
چگونه Multiple sequence alignment خود را تفسیر کنیم؟ ۲۶۰
قسمت های مناسب در یک alignment پروتیینی کدامند؟ ۲۶۰
روی Multiple Sequence Alignment بیشتر کار کنیم ۲۶۱
چگونه توالی های که قابل هم تراز کردن نیستند را مقایسه کنیم؟ ۲۶۵
چگونه با استفاده از Gibbs sampler یک Multiple sequence alignment بسازیم؟ ۲۶۵

فصل دهم: چگونه alignment ها را متناسب با نیازمان تنظیم کرده و برای انتشار عمومی آماده کنیم؟ ۲۶۷

بهترین فرمت، متناسب با نیاز ما، کدام است؟ ۲۶۸
فرمت های اصلی موجود کدامند؟ ۲۷۱
کدام فرمت بهترین انتخاب ماست؟ ۲۷۲
چگونه یک فرمت را به فرمتی دیگر تبدیل کنیم؟ ۲۷۴
مراقب مفقود شدن داده هایتان باشید!!! ۲۷۴
Jalview وسیله ای برای تنظیم Multiple sequence alignment ۲۷۵
شروع کار با Jalview ۲۷۶
چگونه می توانیم به راحتی، دسته ای از توالی ها را به صورت دسته جمعی تنظیم کنیم؟ ۲۷۹
ویژگی های مثبت Jalview ۲۸۲
ذخیره داده ها در Jalview ۲۸۲

چگونه Multiple alignment خود را برای انتشار و گزارش عمومی آماده کنیم؟ .. ۲۸۳
Boxshade چیست؟ ۲۸۳
LOGOS ۲۸۶
برنامه های اینترنتی دیگر برای ویرایش و آنالیز توالی ها ۲۸۷
چند برنامه برای تنظیم و تصحیح alignment ها ۲۸۷
چه سرور هایی در تفسیر Multiple alignment به ما کمک خواهند کرد؟ ۲۸۸
ابزار هایی که برای زیباتر کردن یک multiple sequence alignment می توانند به کار بروند ۲۸۹
بخش چهارم: یک متخصص شوید!! تکنیک های بیوانفورماتیک پیشرفته ۲۹۰
فصل یازدهم: کار کردن با ساختار های سه بعدی پروتئین ها ۲۹۲
از ساختار اولیه تا ساختار ثانویه ۲۹۳
پیش گویی ساختار دوم یک توالی آمینواسیدی ۲۹۳
پیش گویی اطلاعات بیشتر در مورد خصوصیات ساختاری پروتئین ها ۲۹۶
بخش Metaserver در سایت PredictProtein ۲۹۸
یافتن بهترین سرور ۲۹۸
از ساختار اولیه تا ساختار سه بعدی ۲۹۹
یافتن و نمایش یک ساختار سه بعدی ۲۹۹
حدس زدن ساختار سه بعدی یک پروتئین ۳۰۱
بررسی ویژگی های یک توالی در شکل سه بعدی ۳۰۳
بررسی ارتباط بین توالی و ساختار PDB با استفاده از تصاویر متحرک ۳۰۴
آنالیز ارتباط بین توالی و ساختار با استفاده از اشکال متحرک ۳۰۸
فراتر از این فصل ۳۰۹
پیدا کردن پروتئین های با ساختار مشابه ۳۰۹
سایر نمایشگر های PBD ۳۰۹
دسته بندی ساختار های پروتئینی PBD ۳۰۹
مدل سازی از روی یک پروتئین مشابه ۳۱۰
مدل سازی یک پروتئین از روی توالی آمینواسیدی آن در کامپیوتر ۳۱۰
بررسی نحوه تاخوردگی توالی ها و تشکیل ساختار های PDB ۳۱۰
دینامیک و حرکت ساختار های پروتئینی ۳۱۱
پیش گویی برهم کنش ها ۳۱۱

فصل دوازدهم: کار کردن با RNA ۳۱۲
پیش گویی، مدل سازی و طراحی ساختارهای دوم RNA ۳۱۳
استفاده از برنامه mfold ۳۱۴
تفسیر نتایج mfold ۳۱۷
وارد کردن اثر بر هم کنش های RNA در mfold ۳۱۹
جستجوی بانک های اطلاعاتی و ژنوم ها برای یافت توالی های RNA ۳۱۹
یافتن مولکول های tRNA در یک ژنوم ۳۲۰
جستجوی الگوهای RNA با استفاده از PatScan ۳۲۰
یافتن انواع جدید RNA: miRNA ها و siRNA ها ۳۲۳
آنالیز رایگان RNA در اینترنت ۳۲۴
بررسی فرضیه تکامل با استفاده از RNA ریبوزومی ۳۲۴
یافتن RNA های غیرکدکننده ی کوچک ۳۲۵
منابع عمومی RNA ۳۲۶
فصل سیزدهم: طراحی درخت فیلوژنتیک ۳۲۷
کاربردهای درخت فیلوژنتیکی ۳۲۸
آماده کردن اطلاعات فیلوژنیک ۳۲۹
انتخاب توالی مناسب برای یک درخت ۳۲۹
استفاده از توالی های DNA یا پروتئین ۳۳۰
انتخاب توالی ها برای ساخت درخت ژنی یا درخت گونه ها ۳۳۱
ایجاد یک مجموعه کامل ۳۳۳
آماده شدن برای انجام تطابق چند توالی ۳۳۵
انجام تطابق چند توالی ۳۳۵
اطمینان از انتخاب تطابق چند توالی درست ۳۳۵
ساخت درخت مورد نیاز شما ۳۳۷
طراحی درخت مورد نظر شما ۳۳۷
ساختن سریع یک درخت فیلوژنتیک با استفاده از ClustalW ۳۳۸
ساختن یک درخت با استفاده از Phylip ۳۴۱
ساختن درخت بیشترین احتمال با استفاده از PhyML ۳۴۹
اجزاء متشکله یک درخت ۳۵۰

نمایش درخت فیلوژنتیک ۳۵۲
انجام فیلوژنی به صورت رایگان در اینترنت ۳۵۲
یافتن منابع آنلاین ۳۵۲
یافتن منابع عمومی ۳۵۳
مجموعه های ژنهای اورتولوگ ۳۵۴
بخش پنجم: قسمت دهگانه ۳۵۵
فصل چهاردهم: ده (خیلی خوب! دوازده) دستورالعمل برای استفاده از سرور ها ۳۵۷
به خاطر بسپارید: داده های شما بر روی وب هرگز در امان نیستند! ۳۵۸
سرور، بانک اطلاعاتی، و نسخه ی از برنامه های مورد استفاده را به خاطر بسپارید ۳۵۸
شماره شناسایی توالی ها را یادداشت کنید ۳۵۸
ویژگی های برنامه را یادداشت کنید ۳۵۹
نتایج خود را به طور صحیح ذخیره کنید ۳۵۹
استفاده از E-value ۳۶۰
آیا alignment شما قابل اطمینان است؟ ۳۶۰
نتایج سر مرز را با استفاده از برنامه های مختلف چک کنید ۳۶۰
از روش های ثبت نشده اجتناب کنید ۳۶۱
بانک های اطلاعاتی با گذر زمان اعتبار خود را از دست می دهند ۳۶۱
اطلاعاتی که رایگان به نظر می رسند، لزوماً همیشه رایگان نیستند ۳۶۱
چه زمانی برنامه ها را روی کامپیوتر خود نصب کنیم؟ ۳۶۱
فصل پانزدهم: چند منبع مفید در بیوانفورماتیک ۳۶۳
ده بانک اطلاعاتی اصلی ۳۶۴
ده برنامه نرم افزاری اصلی در بیوانفورماتیک ۳۶۴
ده یابنده منبع اصلی ۳۶۴
جاهایی که می توانید جریان های در حال وقوع را دنبال کنید ۳۶۶

مقدمه

در چاپ اول این کتاب، بیوانفورماتیک به عنوان یک رشته علمی نوپا مطرح گردید. در دورانی که ما در آن بسر می بریم، هر فرد با هر نوع گرایش در زمینه بیولوژی، بیوتکنولوژی، پزشکی نوین و یا مهندسی غذا و دارو ، نمی تواند در مقابل این موضوع بی تفاوت بماند. بوسیله این کتاب شما به سرعت و بطور اصولی علوم پایه مرتبط با بیوانفورماتیک را فرا خواهید گرفت. اما کمی صبرکنید!! اگر انتظار یک موضوع پیچیده را دارید باید بگوییم که در اشتباهید. بیوانفورماتیک ، در حقیقت چیزی جز بیولوژی روان و سلیس که در پوششی از کامپیوتر قرار دارد، نمی باشد. بیو انفورماتیک، جستجوی اطلاعات بیولوژیکی، مقایسه توالی ها، نگاهی به ساختمان پروتئین و بطور معمول بیان پرسش هایی در زمینه بیولوژیک و زیست پزشکی[1] بوسیله کامپیوتر می باشد. بیوانفورماتیکی که در این کتاب عرضه می گردد می تواند با صرف کمترین زمان کاری، حاصل ماهها کار شما را بوسیله کامپیوتر نگهداری و آنالیز نماید. در این راه با شرایط استاندارد بیولوژیکی مواجه خواهید شد، پس سعی کنید به معادلات طولانی و اصطلاحات نامفهوم و داده های بدون استفاده در کامپیوتر توجه نکنید. در واقع هدف این کتاب، آموزش راحت و سریع نحوه استفاده از برنامه های بیوانفورماتیک مرتبط با پروژه های شماست. در همه صفحات این کتاب ، راهکارهای لازم برای بهرگیری بیشتر از ابزارهای موجود ارائه شده است. لازم است بدانید که اغلب برنامه ها نیاز به مهارت بالا نداشته و بدون نیاز به نصب بر روی کامپیوتر و تنها به صورت online اجرا خواهند شد. کتاب حاضر به منظور کمک به انجام کارهای شماست . برای بسیاری از امور مهم و استاندارد بیوانفورماتیک که قصد بر عهده گرفتن آنها را دارید ، روش ها و دستورالعمل هایی را یاد خواهید گرفت که به کمک آنها می توانید به سرعت به نتایج مورد نظر دست یابید. برای استفاده از بسیاری از ابزارهایی که در این کتاب شرح داده شده است، نیاز به نصب هیچگونه برنامه ای در کامپیوتر خود ندارید. بسیاری از این برنامه ها از طریق اینترنت و به کمک مرورگر اینترنت (internet browser) قابل اجرا می باشند. اگر می دانید که چه کاری می خواهید انجام دهید و یا حداقل مشخصات کلی از کار خود را می دانید، به سراغ فهرست محتویات کتاب رفته و احتیاجات خود را از طریق این کتاب رفع نمائید. اگر ایده ای در سر دارید ولی نمی توانید آن را به طور دقیق و بوسیله کلمات بیان کنید فصل ۲ ، شما را در این زمینه راهنمایی می کند. در انتهای بسیاری از فصل های کتاب با مفهومی به این قرار آشنا خواهید شد: آنرا به راحتی و بطور مجانی از طریق اینترنت انجام دهید. در این قسمت لیست برخی از سایت های تحت وب که حاوی مفاهیم مشابه با فصل مرتبط می باشند، در اختیارتان قرار خواهد گرفت. این اطلاعات که به دقت انتخاب شده اند را مانند یک چرخ زاپاس تلقی کنید. اگر سایت اصلی اجرا نگردید، لیست مذکور احتمالا جایگزین های دیگری را نیز شامل می شود.

[1] Biomedicine

فرضیات و قوانین

در زمان نوشتن کتاب فرض را بر این گذاشتیم که ...

- ✓ شما کامپیوتر PC با قابلیت اجرای ویندوز مایکروسافت را در اختیار دارید.
- ✓ اتصال به اینترنت را در اختیار دارید. در صورت امکان از نوع پر سرعت ، البته این امر ضروری نمی باشد.
- ✓ با پیش زمینه ای از بیولوژی مولکولی شروع به کار کرده اید. اگر این گونه نیست و یا احتیاج به بررسی بیشتر در این زمینه دارید. فصل ۱ کتاب در واقع مروری سریع بر مطالب پایه بیولوژی مولکولی است.
- ✓ کاربر حرفه ای کامپیوتر نبوده و تنها برای پیشبرد کارهای خود تمایل به استفاده از اینترنت و کامپیوتر را دارید.
- ✓ هدف شما این نیست که یک معلم بیوانفورماتیک باشید، بلکه تنها می خواهید برای حل مشکلات خود، طرز استفاده صحیح ابزار های مورد لزوم را یاد گرفته و وقت خود را برای یادگرفتن اطلاعاتی که لازم نیست، صرف نکنید.
- ✓ بسیاری از شرکت های تخصصی بیوتکنولوژی به این امر واقفند که بیان اطلاعات از طریق اینترنت، از ایمنی خوبی برخوردار نمی باشد. در این جا فرض را بر این قرار می دهیم که اطلاعات مورد آنالیز از طریق اینترنت خیلی هم محرمانه نمی باشند. بسیاری از اطلاعات و سرویس های مذکور در این کتاب نیاز به توافق و داشتن کلمه عبور مربوطه برای وارد شدن است.

شمای کلی از مطالب کتاب

بیو انفورماتیک پهنه وسیعی را شامل می شود که بلندی ها و تپه ها، دره ها و نشانه های فریب دهنده بسیاری در آن به چشم می خورد. برای بیان تمامی اطلاعات و خط مشی های لازم در زمانی کوتاه، آن هم به طور سریع ، مطالب کتاب را در ۵ قسمت تقسیم نموده ایم:

بخش اول : شروع کار با بیوانفورماتیک

اگر کمتر از یک ساعت برای آگاهی از توانایی های بیوانفورماتیک وقت دارید، قسمت اول ، بهترین مکان برای شماست. این قسمت تمام مطالبی را که به منظور کار در زمینه بیوانفورماتیک به آن نیاز پیدا خواهید کرد ، شامل می شود. در این بخش به بیان اطلاعات مفیدی در زمینه بیولوژی مولکولی (هر چند مختصر) و آنالیز توالی ها، خواهیم پرداخت. از طرف دیگر طریقه استفاده از ابزار های اصلی بیوانفورماتیک را نشان خواهیم داد.

بخش دوم: راهنمایی ضروری در خصوص بیوانفورماتیک

در صورت تمایل به یافتن تمامی اطلاعات انتشار پیدا کرده در مورد توالی مورد نظرتان قسمت دوم کتاب می تواند راهگشای خوبی باشد. این قسمت راهنمای مناسبی برای دست و پنجه نرم کردن با مسائل پیشرو در زمینه بیوانفورماتیک است : توالی های DNA و پروتئین ها و اطلاعات پیرامون آنها. از طرف دیگر چگونگی یافتن توالی های در دسترس از میان میلیون ها داده را به شما آموزش خواهیم داد و اینکه چگونه بتوانید تمامی اطلاعاتی را که در جهان راجع به توالی اختصاصی شما انتشار یافته است (دست کم آنهایی را که بصورت online در دسترس هستند)، پیدا نموده و در کنار هم قراردهید، این مسئله شما را هیجان زده خواهد کرد!

بخش سوم : در آنالیز توالی ها مهارت کافی پیدا نمائید!

اگر قصد مقایسه توالی ها را دارید، این قسمت مربوط به شماست. در این بخش نحوه جستجوی اطلاعات پیرامون توالی های مشابه و نیز روش مقایسه دو یا چند توالی را به شما نشان خواهیم داد. همچنین، نحوه جمع آوری اطلاعات با جزئیات بیشتر پیرامون عملکرد یک ژن را از طریق مقایسه توالی ها آموزش خواهیم داد و سرانجام ، نکاتی راجع به چگونگی تولید، ویرایش و زیباسازی اطلاعات ژنتیکی جمع آوری شده ، ارائه خواهیم داد تا بتوانید از این راهکارها به هنگام انتشار استفاده نمائید.

بخش چهارم: یک متخصص شوید!! تکنیک های بیوانفورماتیک پیشرفته

برای بهره برداری بیشتر از این قسمت ، در ابتدا باید ایده ای مناسب پیرامون آنچه که در پی آن هستید، داشته باشید. سوالاتی چون :

- چگونه ساختار یک پروتئین را پیش بینی کنید؟
- چگونه ساختار یک RNA را پیش بینی کنید؟
- چگونه یک آنالیز مرتبط با تکامل نژادی را پایه گذاری نمایید؟

این امور هر چند ممکن است در ابتدای کار کمی پیچیده به نظر برسند ولی در عمل به راحتی و بوسیله یک کامپیوتر PC ساده و همچنین منابع اینترنتی شرح داده شده در این قسمت ، قابل انجام خواهند بود.

بخش پنجم : قسمت دهگانه

به بازار ما خوش آمدید!! اگر نتوانستید آنچه را که می خواهید در قسمت های دیگر جستجو کرده و بیابید، به مکان مناسبی آمده اید. منابع بسیار زیادی، آنهم به صورت online، پیرامون بیوانفورماتیک وجود دارد. با این حال جستجوی آنها اغلب طاقت فرسا و خسته کننده است. برای بسیاری از دانشجویان و اطرافیان آنها شاید گزارش online نیمسال تحصیلی و پیدا نمودن تمامی اطلاعات مورد نظر با دقت بالا آنهم تنها با یک کلید واژه ساده ، امری ترسناک جلوه نماید. در قسمت دهگانه لیستی از منابع اصلی مورد نیاز را به عنوان نقطه شروع در اختیار شما قرار می دهیم. شانس شما اینست که برنامه یا سروری[1] که به دنبال آن هستید به

[1] server

اندازه یک یا دو فشار دکمه موس[1] آنطرف تر نیست. به منظور اطمینان از ایمن بودن کارهای آزمایشگاهی خود تحت پوشش اینترنت، در این قسمت لیستی ارائه می گردد که شامل ده جزء مهم است.

علامت های استفاده شده در این کتاب

در این کتاب برای نشان دادن مطالب مهم، از علامت هایی استفاده شده است:

بعضی نکات تکنیکی خاص را نشان می دهد. اگر این نکات را نخوانید، اتفاق مهمی نمی افتد، اما برای اینکه کاملاً به مطالبی که می خوانید مسلط شوید، خواندن این قسمت ها می تواند کمک کننده باشد.

این علامت یک میانبر، راه ساده تر یا زیرکانه تر را نشان می دهد. به طور کلی می توان گفت با رعایت این نکات می توانید در وقت خود صرفه جویی کنید.

هنگام استفاده از اینترنت دام های زیادی سر راه شما وجود دارد. این علامت این وجود این دام ها در مسیر شما را هشدار می دهد. این علامت را جدی بگیرید، بخصوص هنگامی که دستورالعمل های چند مرحله ای این کتاب را انجام می دهید.

این علامت به شما توصیه می کند که مطلبی را به خاطر بسپارید. این مطالب می توانند یکی از معدود نکات مهمی باشند که همیشه باید در ذهن شما وجود داشته باشند.

هدف نهایی این کتاب

اگر درباره بیوانفورماتیک اطلاعات کافی ندارید، این کتاب می تواند قوت قلبی برای شما باشد. بیوانفورماتیک از آنچه تصور می کنید بسیار ساده تر است. برای بسیاری از تازه کارها ، مسئله اساسی یافتن پاسخ انبوهی از سوالات است که بوسیله چنین ابزار های جدیدی می توان به آنها پاسخ داد. اگر یک زیست شناس هستید، اجازه ندهید که کامپیوتر ترسی در دلتان وارد کند. بدانید بیوانفورماتیک در حقیقت چیزی جز بیولوژی منظم و روان نیست که در پشت کامپیوتر مخفی شده است. جالب تر اینکه به سادگی با اتصال به اینترنت می توانید داده هایی را که شامل همه اطلاعات پیرامون بیولوژی انسانی است، جستجو کنید. شما تمامی این کارها را با ابزار های پیچیده و حرفه ای که توسط انسان ها ساخته و در دسترس قرار گرفته اند ، به انجام می رسانید و این کار چقدر خرج دارد؟ در حقیقت هیچی! اگر پروژه شما در زمینه بیولوژی مولکولی است، این حالت معادل داشتن یک آزمایشگاه با همه تجهیزات مدرن و گرانبها است که هر وقت اراده نمایید اطلاعات ذخیره شده را در اختیارتان قرار می دهد و تنها تفاوت موجود این است که با وجود تلاش های بسیار هیچگاه در به آتش کشیدن این آزمایشگاه موفق نخواهید بود! حتما تاکنون متوجه شده اید که

[1] mouse

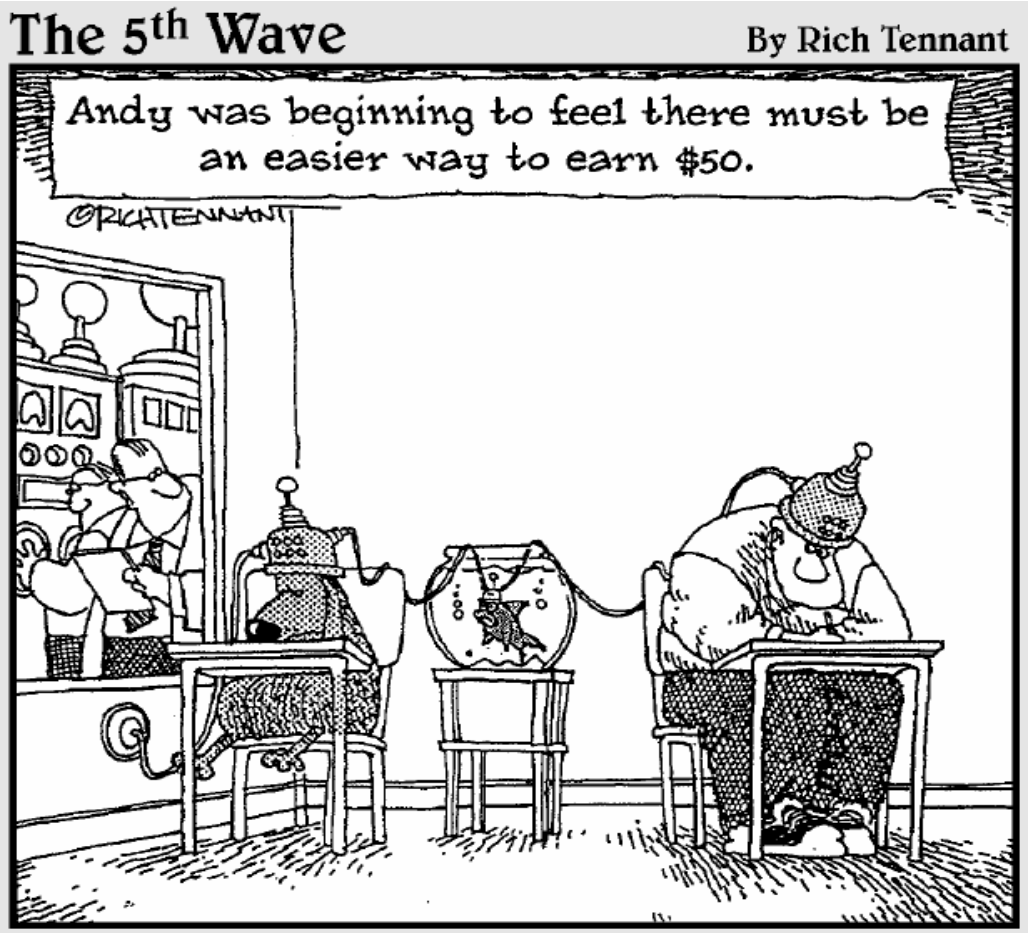
در حقیقت همه چیز به سادگی یک یا دو کلیک برروی موس است. شبکه جهانی اینترنت حد و مرز مشخصی ندارد و به خاطر گستردگی فراوانی که دارد کاملا بی رنگ و بی نشان است. با این حال اگر شما از سرزمین ثروتمند یا فقیری هستید و یا اگر دانش آموز سال اول و یا یک دانشمند یا برنده جایزه نوبل هستید، به طور یکسان به اطلاعات با کیفیت بالا دسترسی خواهید داشت. هیچ روش علمی دیگری تاکنون اینگونه به صورت گسترده و همگانی وجود نداشته است. این کتاب در حقیقت یک کتاب درسی و یا یک رساله نیست بلکه یک کتاب آشپزی است! و ما به آن افتخار می کنیم. این کتاب شامل دستورالعمل هایی است که یا نتیجه سالها تلاش دانشجویان است و یا ما خود آنها را یافته ایم. نحوه بکارگیری و تطبیق اطلاعات بیولوژیکی یک کار شخصی است، اما ما به شما این اطمینان را می دهیم که به تدریج تمامی راههای موجود برای انجام این قبیل کارها را فراخواهید گرفت. اگر می خواهید سریع تر به نتیجه برسید می توانید از بعضی از راه حلهای در دسترسی که برای شما تدارک دیده ایم بهره بگیرید. هیچ روش علمی به اندازه بیولوژی در این دهکده جهانی سودمند نمی باشد. هر سوالی که دارید و هر آنچه می خواهید انجام دهید ابتدا از اینترنت شروع نمائید، این بهترین کار ممکن است. با این وجود همواره در کنار بهترین ها، بدترین هایی نیز وجود دارد و دنیای مجازی وOnline نیز از این قاعده مستثنی نمی باشد. بنابراین در این مورد هم مانند دنیای واقعی رفتار کرده و تنها به موسسات و سایتهایی که به خوبی می شناسید، اعتماد کنید.

این کتاب تا جایی که در توان ما بود به روز می باشد ولی در جهان کنونی حتی بعد از اتمام کار، جمع آوری مدارک و انتشار اطلاعات برای عموم مردم از طریق کتاب حاضر هم در یک جا ثابت نمانده و دائم در حال تغییر است! برای بعضی از شما که تمایل به بروز رسانی اطلاعات در زمینه های قابل رشد بیوانفورماتیک دارید (که شامل لیستی از لینک های مورد علاقه ما در زمینه بیوانفورماتیک می باشد) و نمی خواهید منتظر ویرایش های بعدی کتاب باشید به سایت : www.dummies.com رجوع کنید.

بعضی مواقع با اتصال به اینترنت ومشاهده اینکه تمامی کارها توسط دیگران انجام می گیرد، احساس افسردگی می کنید. در دنیای امروز که تمامی جهان در ارتباط نزدیک با هم به سر می برند، روشن است که با سوالهای محدوی مواجه شوید. این خبر بدی است. خبر خوب اینکه برای پرسشهای موجود، جواب های بسیار زیادی وجود دارد. هیچگاه این فرض را نادیده نگیرید که پاسخ شما ممکن است در جهان بهترین باشد، حتی برای چند روز!!

بخش اول

شروع کار در بیوانفورماتیک



ترجمه: سجاد عباسی

ویرایش: حسن کربکندی

در این بخش می خوانیم...

بیوانفورماتیک رشته ای علمی و جدید است و هیچ کس نباید از اینکه ایده ای در مورد این رشته مهیج ندارد احساس خجالت کند. نگران نباشید، بعد از تمام کردن این کتاب شما می توانید به بهترین نحو به زبان بیوانفورماتیک صحبت کنید.

این فصل را با یک یادآوری سریع از مطالبی که باید در مورد DNA و پروتئین ها بدانید شروع می کنیم تا برداشت صحیحی از این کتاب داشته باشید. همچنین یک نگرش کلی در مورد ابزارهای اصلی بیوانفورماتیک که در اینترنت وجود دارند به شما می دهیم.

ما زیاد وارد جزئیات نمی شویم اما اگر می خواهید بدانید برای کار کردن در زمینه بیوانفورماتیک به چه صفحه ای از اینترنت باید بروید یا چه دکمه ای را باید فشار دهید با ما همراه باشید چون این کتاب درست همان چیزی است که شما می خواهید.

فصل اول

آنچه بیوانفورماتیک می تواند برای ما انجام دهد

در این فصل می خوانیم :

- ✓ تعریف بیوانفورماتیک
- ✓ درک ارتباط بین بیولوژی نوین، ژنومیک و بیوانفورماتیک
- ✓ تعیین آنکه با کمک بیوانفورماتیک به چه سؤال های بیولوژیکی می توانید بسرعت جواب دهید.

.....

" شیمی آلی شیمی ترکیبات کربن است. بیوشیمی، مطالعه ترکیباتی از کربن است که می خزند."

— Mike Adam

به نظر می رسد که زیست شناسان در حال تدوین لغت نامه ای از لغاتی هستند که با بیو- شروع می شوند : بیوشیمی ، بیومتریک (زیست سنجی)، بیوفیزیک، بیوتکنولوژی، بیوهازارد و حتی بیوتروریسم. اما ببینیم از لغت جدیدی که وارد این لغت نامه شده، یعنی بیوانفورماتیک ، چه خبر؟

بیوانفورماتیک چیست؟

در دنیای امروز، از بیوانفورماتیک همان قدر که در بانکداری و کنترل هواپیما استفاده می شود، در زیست شناسی نیز مورد مصرف قرار می گیرد. بسیاری از کارهایی که در این حرفه ها با کامپیوتر انجام می شود،کارهای معمولی است که اکثر ما انجام می دهیم. ما همگی تمایل داریم مقدار زیادی یادداشت های یادآوری بنویسیم و تعداد زیادی نامه الکترونیکی بفرستیم. بسیاری از ما از صفحات گسترده[1] مثل Excel استفاده می کنیم، و مقادیر زیادی اطلاعات که هرگز دوباره بازبینی نمی شوند را در سیستم های بایگانی پیچیده ذخیره می کنیم.

[1] Spreadsheets

اما در کنار این کارهای عمومی، زیست شناسان از کامپیوتر همچنین برای حل مشکلاتی که اختصاصاً به زیست شناسی مربوط می شود و به هیچ وجه مورد علاقه یک کارمند بانک یا یک کارمند مراقبت پرواز نیست، استفاده می کنند. این کارهای اختصاصی مربوط به زیست شناسی وقتی یک جا جمع شوند رشته بیوانفورماتیک را می سازند. به طور اختصاصی تر بیوانفورماتیک شاخه محاسبات کامپیوتری رشته زیست شناسی مولکولی است.

اکنون زمان آن است که تاریخچه مختصری از بیوانفورماتیک را ذکر کنیم. قبل از به وجود آمدن بیوانفورماتیک فقط دو راه برای انجام آزمایشات زیست شناسی وجود داشت: درون یک ارگانیسم زنده (که in vivo نامیده شد) و در یک محیط مصنوعی (که in vitro نامیده شد ، برگرفته از عبارت لاتین " درشیشه"). به همین ترتیب اگر بخواهیم یک معادل هم برای بیوانفورماتیک پیدا کنیم، "زیست شناسی in silico" اصطلاح مناسبی است. این واژه از تراشه های سیلیکون که ریز پردازشگر ها را می سازند، برگرفته شده است.

این روش انجام آزمایشات بیولوژی خیلی رایج شده است، اما این رواج به معنی بیهوده و کم ارزش بودن آن نیست. در حقیقت بیوانفورماتیک نقطه مرکزی تحولات اخیر در زیست شناسی است. از این تحولات می توان موارد زیر را نام برد :

رمز گشایی ژنوم انسان ، " system biology" (در نظر گرفتن یک تصویر کلی از یک سیستم)، زیست فناوری های جدید، تکنیک های جدید پزشکی قانونی، و همچنین داروهای اختصاصی برای هر شخص در آینده.

به دلیل نقش کلیدی بیوانفورماتیک در تحولات مهم ایجاد شده در زیست شناسی مولکولی، افراد در زمینه های مختلفی از علوم با بیوانفورماتیک سروکار دارند. اگر یک دانشجوی زیست شناسی، پزشکی یا علوم کامپیوتریا متخصص صنایع داروسازی هستید یا یک وکیل یا یک پلیس که با آزمایش DNA سروکار دارد، مصرف کننده ای که نگران ارگانیسم های دست ورزی ژنتیکی شده[1] است ، یا حتی سرمایه گذار بورس NASDAQ که علاقمند به تأسیس یک شرکت است، با بیوانفورماتیک برخورد خواهید داشت. برای این که یک درک سریع از بیوانفورماتیک پیدا کنیم، بهتر است مستقیماً وارد دریای بیوانفورماتیک شویم. برای شروع کار، بعضی از مفاهیم اصولی زیست شناسی مولکولی و سؤالاتی که در این رابطه به ذهن خطور می کند را مرور می کنیم . همه این مطالب روی هم تعریف بیوانفورماتیک خواهد بود.

آنالیز توالی های پروتئینی

وقتی گوشت استیک می خورید فوراً متوجه پروتئین هامی شوید. جوانه های چشایی شما فوراً پروتئین ها را می شناسند، هر چند عقل استدلال گر شما آنقدر مشغول باشد که نتواند این شناخت را ایجاد کند. شاید برای شما جالب باشد که بدانید پروتئین ها به فراوانی در گوشت ماهی و گیاهان هم یافت می شوند. علاوه بر این ، همه این پروتئین ها ازواحدهای سازنده یکسانی تشکیل شده اند که اسیدهای آمینه نام دارند. اسیدهای آمینه هم به نوبه خود مولکول های آلی بسیار پیچیده ای متشکل از اتم های کربن، هیدروژن، اکسیژن و گوگرد هستند. بنابراین ساختار کلی پروتئین ها (این بار ساختاری که عقل استدلال گر شما می تواند آن را درک کند، اما جوانه های چشایی شما قادر به این کار نیست)، چیزی شبیه $C_{1200}H_{2400}O_{600}N_{300}S_{100}$ است.

[1] Genetically Modified Organisms or GMO

پس از پیدایش رشته بیوشیمی، اولین تلاش ها در این رشته به پیدا کردن راه بهتری برای نامگذاری پروتئین ها اختصاص داده شد، نامی که ترجیحاً به شکل یک فرمول و نشان دهنده خواص بیولوژیکی و یا حتی تغذیه ای پروتئین ها باشد. با گذشت زمان متخصصین بیوشیمی دریافتند که فقط ۲۰ نوع اسید آمینه به نام هایی مثل آلانین، گلیسین، تیروزین، گلوتامین و... وجود دارد و پروتئین ها مولکول های بسیار بزرگی(ماکرومولکول) هستندکه ازکنار هم قرار گرفتن تعداد زیادی از همین ۲۰ آمینو اسید (معمولاً۱۰۰-۵۰۰ عدد) تشکیل شده اند. جدول ۱-۱ نام کامل ، کد سه حرفی و کد یک حرفی این ۲۰ اسید آمینه را فهرست کرده است. کد یک حرفی،کدی است که " کد آیوپاک"[1] هم نامیده می شود.

بعداً مشخص شد که تعداد کل اسیدهای آمینه و نسبت آنها در یک نوع پروتئین (مثلاً انسولین یا میوگلوبین) همواره ثابت است. پس می توان فرمول بهتری برای یک پروتئین درنظر گرفت مثلاً : انسولین = [۳۰ (گلیسین) + ۴۴ (آلانین) + ۵ (تیروزین) + ۱۴ (گلوتامین) +...] و سرانجام متخصصین بیوشیمی دریافتند که این اسیدهای آمینه به هم متصل شده و به شکل یک زنجیر در می آیند، بنابراین خصوصیات واقعی یک پروتئین نه تنها به نوع اسیدهای آمینه، بلکه به ترتیب قرار گرفتن آنها در کنارهم نیز بستگی دارد.

#	1-Letter Code	3-Letter Code	Name
1	A	Ala	Alanine
2	R	Arg	Arginine
3	N	Asn	Asparagine
4	D	Asp	Aspartic acid
5	C	Cys	Cysteine
6	Q	Gln	Glutamine
7	E	Glu	Glutamic acid
8	G	Gly	Glycine
9	H	His	Histidine
10	I	Ile	Isoleucine
11	L	Leu	Leucine
12	K	Lys	Lysine
13	M	Met	Methionine
14	F	Phe	Phenylalanine
15	P	Pro	Proline
16	S	Ser	Serine
17	T	Thr	Threonine
18	W	Trp	Tryptophan
19	Y	Tyr	Tyrosine
20	V	Val	Valine

جدول ۱-۱: بیست اسید آمینه و کدهای رسمی آنها

اولین بار توالی آمینواسیدی انسولین در سال 1952 مشخص شد. ساختار واقعی مولکول انسولین انسانی که مسئول همه خصوصیات بیولوژیکی آن است، یک زنجیر متشکل از ۱۱۰ اسید آمینه به صورت زیر است:

Insulin = MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAE
DLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

[1] IUPAC

اکنون که بیش از ۵۰ سال از آن زمان می گذرد هنوز آنالیز توالی های پروتئینی بحث اصلی بیوانفورماتیک در آزمایشگاه های جهان است (برای آشنایی با چگونگی آنالیز توالی های پروتئینی به فصل های ۲ ،۴،۶ تا ۱۱ مراجعه کنید).

یک تاریخچه مختصر از آنالیز توالی

یافتن توالی انسولین گذشته از این که سبب شد Alfred Sanger اولین جایزه نوبل خود رادریافت کند، باعث به وجود آمدن دوران جدیدی در زیست شناسی ساختاری و مولکولی شد. در اوایل دهه ۱۹۶۰ ،توالی های پروتئینی مشخص شده، زیاد و انبوه شدند و چون کامپیوتر هایی که قادر به پردازش آنها باشند هنوز توسعه پیدا نکرده بودند، اطلاعات مربوط به توالی ها به صورت دستی جمع آوری و آنالیز می شدند؛ بدین ترتیب که این اطلاعات را روی کاغذ نوشته کنار هم روی دیوار آزمایشگاه ها می چسباندند و با جابجا کردن آنها نهایتاً به ترتیب مطلوب می رسیدند (این فرآیند الانPattern matching نامیده می شود.).

به محض اینکه اولین کامپیوتر ها (به بزرگی و با سرعت یک لوکوموتیو و با حافظه ۸ کیلو بایتی!) به وجود آمدند ، تعدادی از زیست شناسان وارد کردن این الگوریتم های دستی به کامپیوترها را آغاز کردند. این کار کاملاً نو بود و هیچ کس تا آن زمان آن را انجام نداده بود. اکثر این روش ها از هیچ اختراع شدند و با پیشرفت تدریجی آنها یک حوزه جدید در تحقیقات علمی یعنی "آنالیز توالی های پروتئینی به وسیله کامپیوتر "به وجود آمد و منشأ پیدایش رشته بیوانفورماتیک شد.

هفت کد آمینو اسید دیگر

وقتی با بانک های اطلاعاتی یا برنامه های آنالیز کار می کنید گاهی اوقات ممکن است حروف غیر معمولی درتوالی پروتئین شما ظاهر شوند. این حروف یا مربوط به آمینواسیدهای غیر معمول هستند یا درجات مختلفی از ابهام در جاهای مختلفی توالی پروتئینی را نشان می دهند (که به دلیل فقدان اطلاعات کافی است).

این حروف در جدول زیر فهرست شده اند.

جدول ۲-۱: هفت کد مربوط به آمینو اسیدهای مبهم یا استثنایی

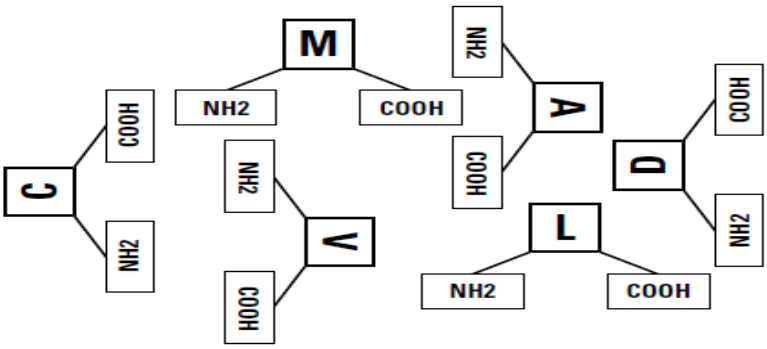
1-Letter Code	*3-Letter Code*	*Meaning*
B	Asn or Asp	Asparagine or aspartic acid
J	Xle	Isoleucine or leucine
O (letter)	Pyl	Pyrrolysine
U	Sec	Selenocysteine
Z	Gln or Glu	Glutamine or glutamic acid
X	Xaa	Any residue
--	-----	No corresponding residue (gap)

کدهای B , Z (که الان منسوخ شده اند) نشان می دهند که در روزهای ابتدایی تعیین توالی پروتئین ،تفکیک آسپارتات از آسپارژین و یا گلوتحامات از گلوتامین چقدر مشکل بوده است. و کد J نشان دهنده دشواری تفکیک لوسین از ایزولوسیون با استفاده از آخرین تکنیک تعیین توالی یعنی اسپکترومتری جرمی است. آمینو اسیدهای مبهم Sec , Pyl به ترتیب مربوط به کدون های توقف[1] UAG , UGA در فرآیند ترجمه هستند . کد α که هنوز به فراوانی مورد استفاده قرار می گیرد مربوط به موقعیت هایی از توالی پروتئینی است که آمینواسید مربوط به آن مشخص نشده است. همچنین علامت " ~ " در برنامه های تعیین توالی نشان دهنده محل هایی از توالی پروتئین است که به نظر می رسد تشخیص داده نشده اند.

خواندن توالی پروتئینی از N تا C

ساختار مولکولی ۲۰ آمینو اسیدی که در پروتئین ها یافت می شود با هم متفاوت اند اما همه آنها یک جفت "چنگک" مشابه دارند : NH2– , COOH – . این چنگک ها یا گروه های عاملی باعث تشکیل پیوندهای پپتیدی بین آمینواسیدهای متوالی می شوند. شکل ۱-۱ چند آمینواسید آزاد و چنگک های آن را نشان می دهد.

مولکول پروتئین زمانی تشکیل می شود که یک گروه NH_2– آزاد به یک گروه COOH– آزاد متصل شده پیوند پپتیدی CO –NH تشکیل دهد . شکل ۲-۱ حالت شماتیک زنجیر حاصل را نشان می دهد. زنجیر پروتئین پس از تشکیل در یک انتها دارای گروه NH2– آزاد و در انتهای دیگر دارای گروه COOH– آزاد است که به ترتیب "انتهای آمینی" [2] و "انتهای کربوکسیلی" [3] نامیده می شوند.

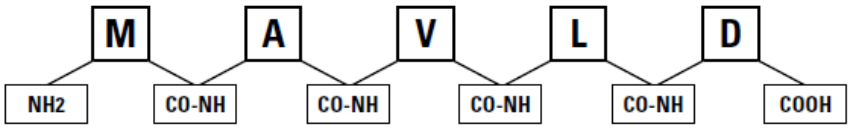


شکل ۱-۱: آمینواسید های آزاد

دانستن این مطلب مهم است زیرا براساس این قرارداد در کتاب ها بانک های اطلاعاتی و ... توالی یک پروتئین (یا قطعه ای از یک پروتئین) به صورت توالی آمینواسیدها که از انتهای آمینی به کربوکسیلی ردیف شده اند، نشان داده می شود. بنابراین توالی پروتئین نمایشی کوتاه (!) ، به این صورت خواهد شد:

MAVLD = Met- Ala- Val- Leu- Asp =متیونین – آلانین – والین– لوسین– آسپارتیک اسید

[1] Stop Codons
[2] N-terminus
[3] C-terminus



شکل ۲-۱ : آمینواسیدها که به صورت زنجیروار در کنارهم قرار گرفته و یک مولکول پروتئین تشکیل داده اند.

کارکردن با ساختارهای سه بعدی پروتئین ها

صرفا با داشتن توالی دقیق پروتئین، همه خصوصیات بیولوژیکی آن (مثلاً توانایی آن درهضم قند یا توانایی قرارگرفتن آن در ساختمان فیبر ماهیچه ای) را نمی توان توجیه کرد. این خصوصیات از شکل سه بعدی (3-D) که توالی آمینواسیدی به خود می گیرد ناشی می شوند. یک مولکول پروتئین بلافاصله پس از ساخته شدن یک شئ زنجیرمانند و بسیار انعطاف پذیر نیست بلکه بیشتر شبیه یک توپ متراکم است که از روی هم جمع شدن رشته آمینواسیدی حاصل شده است. توالی آمینو اسیدی پروتئین اختصاصاً شکل سه بعدی نهایی آن را تعیین می کند چون بعضی انواع آمینواسیدی (مثلاً آمینو اسیدهای چربی دوست J,V,L) به هیچ وجه تمایل ندارند در سطح مولکول و در تماس با آب قرار گیرند. در حالی که دیگر آمینو اسیدها (مثلاً آمینواسیدهای آب دوست K,S,D) فعالانه تمایل دارند در تماس با آب و برهم کنش داشته باشند. از عوامل دیگری که برشکل سه بعدی مولکول پروتئین اثر می گذارد می توان به شارژالکتریکی بعضی آمینواسیدها و توانایی آمینواسیدهای مجاور در جفت شدن با هم اشاره کرد. اولین بارد سال 1958 دکتر Kendrew و دکتر Perutz شکل سه بعدی یک پروتئین را با استفاده از تکنیک پیچیده کریستالوگرافی با اشعه x تعیین کردند . این شاهکار بزرگ، گذشته از این که سبب شد یک جایزه نوبل دیگر هم به رشته در حال رشد زیست شناسی مولکولی اختصاص داده شود، باعث کشف این موضوع شد که پروتئین ها اشکال دقیق و اختصاصی متشکل از توالی آمینواسیدها هستند. از این رو نتیجه گرفتند که پروتئین های با توالی آمینواسیدهای مشابه شکل سه بعدی مشابهی خواهن داشت و برعکس این مطلب هم صادق است، یعنی پروتئین هایی با ساختار مشابه از آمینواسیدهای مشابهی تشکیل شده اند. همچنین مشخص شد که عملکرد یک پروتئین هم مستقیماً به ساختار سه بعدی آن مربوط می شود.

وجود این ارتباط منطقی یعنی : عملکرد ← ساختار توالی ←آمینواسیدها، اثبات شد و امروزه این ارتباط یک اصل مهم در بیولوژی و بیوانفورماتیک است. بدیهی است که کار کردن با مدل های ساختار های پروتئینی به وسیله کامپیوتر بسیارآسان تر از کارکردن با هزاران مدل پلاستیکی از ساختار سه بعدی پروتئین ها است. در نتیجه هم اکنون بخش بزرگ و فزاینده ای از تلاش ها در زمینه بیو انفورماتیک به تولید یک ابزار کامپیوتری که بتواند بین توالی آمینواسیدی و شکل سه بعدی پروتئین ارتباط ایجاد کند اختصاص داده شده است. این بخش از ریشه بیوانفورماتیک Structural Bioinformatics نامیده می شوند. با وجود منابع رایگان فراوان موجود در اینترنت شما به آسانی می توانید چند ساختار پروتئینی زیبا را به کامپیوتر خود منتقل کرده، بازی کردن با این ساختارها را شروع کنید (اگر می خواهید بدانید چه طور اینکار را انجام دهید فصل ۱۱ را مطالعه کنید). اما قبل از این که فصل ۱۱ را بخوانید شکل ۳-۱ می تواند به شما نشان دهد که ساختار سه بعدی یک پروتئین متشکل از ۴۰۰ اسیدآمینه در یک مانیتور سیاه وسفید چگونه به نظر خواهد رسید.

فراموش نکنید؛ مولکول های پروتئینی با وجود پیچیدگی شگفت انگیزشان خیلی کوچک هستند. مثلاً یک جعبه با اضلاع ۷۰/۱۰۰۰۰۰۰ میلی متری برای جای دادن مولکول پروتئینی شکل ۳-۱ کافی است . هزاران نوع مولکول پروتئینی و از هر نوع هزاران نسخه در یک سلول باکتریایی وجود دارد. همین یک گواه کافی است بر این که "حیات موجودات ساده نیست!"

شکل۳-۱: مثالی از ساختار سه بعدی پروتئین

سرفصل های مرتبط با بیوانفورماتیک در این کتاب

مطالعه توالی های پروتئینی می تواند کاملاً پیچیده شود، آنقدر پیچیده که برای پوشش دادن همه جنبه های آن یک کتاب بسیار قطور نیاز است. اما ما کمی اختصاصی تر عمل کردیم و روی جنبه هایی از توالی یابی از پروتئینی متمرکز شده ایم که بیوانفورماتیک در آنها بیشترین استفاده را دارد. سرفصل هایی که اختصاصاً با توالی یابی پروتئین مرتبط هستند و فصولی از کتاب که در مورد جزئیات آنها توضیح داده در زیر فهرست شده اند:

- ✓ بازیابی توالی های پروتئینی از بانک های اطلاعاتی (فصل های۲، ۳ و ۴)
- ✓ محاسبه اجزای سازنده، وزن مولکولی، نقطه ایزوالکتریک ودیگر پارامترهای آمینواسید (فصل ۶)
- ✓ محاسبه میزان آبدوست یا آبگریز بودن پروتئین، پیش بینی محل های آنتی ژنی ممکن، تعییین محل قطعاتی از پروتئین که روی غشا پل می زنند. (فصل ۶)
- ✓ پیشگویی عوامل اصلی ایجادکننده ساختمان دوم (فصل ۱۱و ۶)
- ✓ پیشگویی تشکیلات ناحیه ای پروتئین ها (فصل های ۱۱-۹-۷-۶)
- ✓ تجسم ساختار سه بعدی پروتئین ها (فصل ۱۱)
- ✓ پیشگویی ساختار سه بعدیپروتئین ها از روی توالی آنها (فصل ۱۱)
- ✓ یافتن همه پروتئین هایی که توالی مشابه دارند. (فصل ۷)
- ✓ دسته بندی کردن پروتئین ها در چند خانواده (فصل ۹-۸-۷)

- ✓ یافتن بهترین تطابق بین دو یا تعداد بیشتری پروتئین (فصل ۹-۸)
- ✓ یافتن روابط تکاملی بین پروتئین ها، پروتئین های طراح و شجره خانوادگی (فصل۷، ۹، ۱۱ و ۱۳)

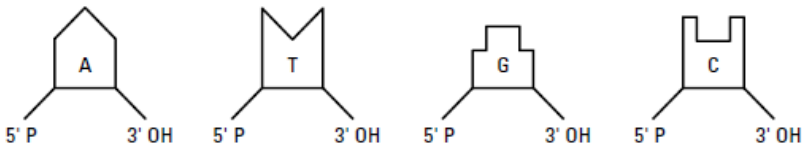
آنالیز توالی های DNA

در دهه ۱۹۵۰، زمانی که دانشمندانی چون Kendrew و Perutz در تلاش برای یافتن ساختار سه بعدی یک پروتئین برای اولین بار بودند ، دیگر زیست شناسان با آزمایشات ژنتیکی بسیار پیچیده شواهد زیادی بدست آورده بودند دال بر این که ماده سازنده ژن های ما یعنی دئوکسی ریبونوکلئیک اسید (DNA) یک ماکرو مولکول زنجیر مانند دراز است که به صورت یک مارپیچ دوگانه حول یک محور پیچ خورده است. آنها همچنین می دانستند که اجزای سازند مولکول DNA نوکلئوتید نام دارند و پیوند بین نوکلئوتید های متقابل است که باعث می شود دو زنجیره در کناره هم نگه داشته شوند.

اما به هر حال زیست شناسان باید تا سال ها بعد (دهه ۱۹۶۰) صبر می کردند تا بتوانند مولکول های DNA را توالی یابی کنند و به توالی نوکلئوتیدی ژن ها دستیابی پیدا کنند. این واقعه یک انقلاب در زیست شناسی بود (باعث شد A.Songer دومین جایزه نوبل خودش را بگیرد!)، چون الفبای ۴ حرفی توالی DNA در مقایسه باا لفبای ۲۰ حرفی توالی پروتئین باعث می شدکه خواندن آنها راحت تر باشد و سریع تر به سمت کامپیوتری شدن پیش برود. به همین دلیل در دنیای امروز روند توالی یابی DNA سریع تر از تولی یابی پروتئین است.

خواندن صحیح توالی های DNA

مشابه آن چه در مورد ۲۰ آمینواسید سازنده پروتئین ها دیده شد. ۴ نوع نوکلئوتید سازنده DNA هم بدنه های متفاوتی دارند اما همه آنها دارای یک جفت چنگک مشابه هستند: ′5 فسفوریل و′3 هیدروکسیل. اعداد′5 و′3 نشان دهنده موقعیت گروه های فسفوریل و هیدروکسیل روی مولکول قنددئوکسی ریبوز است. شکل ۴-۱، چهار نوع نوکلئوتید را به صورت آزاد نشان می دهد.

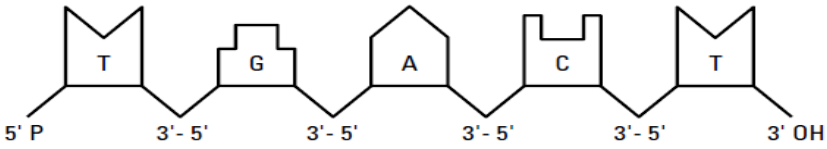


شکل ۴-۱: چهار نوع نوکلئوتید سازنده DNA

مولکول DNA از طریق تشکیل پیوند بین موقعیت های5 و′3 نوکلئوتیدها حاصل می شود. شکل ۵-۱ رشته DNA حاصل ازاتصال نوکلئوتیدها به هم را نشان می دهد. پس از این که نوکلئوتیدها به هم متصل شدند رشته DNA دارای یک گروه فسفوریل آزاد (Po4) در انتهای 5 خود و یک گروه هیدروکسیل آزاد (OH) در انتهای ′3 است.

براساس قرارداد در کتاب ها ، بانک های اطلاعاتی ، مقالات و برنامه ها ، یک توالی DNA همواره از انتهای 5′ به 3′ نوشته می شود. مثلاً توالی رشته کوتاه ! DNA که در شکل ۵-۱ نشان داده شد به این صورت است:

تیمین – سیتوزین- آدنین – گوانین- تیمین = TGACT



شکل ۵-۱: نوکلئوتید های تشکیل دهنده مولکول DNA به صورت رشته ای

دو شکل خواندن توالی DNA

در سال ۱۹۵۳ Watson و Crick (در همان آزمایشگاهی که Kendrew و Perutz در تلاش برای پیدا کردن ساختار سه بعدی پروتئین برای اولین بار بودند) ساختار معروف مارپیچ مضاعف مولکول DNA را کشف کردند . امروزه هم از این نردبان مارپیچی یک ذهنیت دارند؛ شاید ظرافت و زیبایی مارپیچ دوگانه DNA باعث شده کهاین ساختار معروف ترین مفهوم در زیست شناسی مولکولی باشد . اهمیت این کشف بزرگ (حاصل شدن یک جایزه نوبل دیگر برای رشته زیست شناسی مولکولی!) به خاطر شکل مارپیچی مولکول DNA نبود بلکه به خاطر روشن شدن این واقعیت بود که مولکول DNA از دو رشته مکمل تشکیل شده است. شکل ۶-۱ این مطلب را به خوبی نشان می دهد.

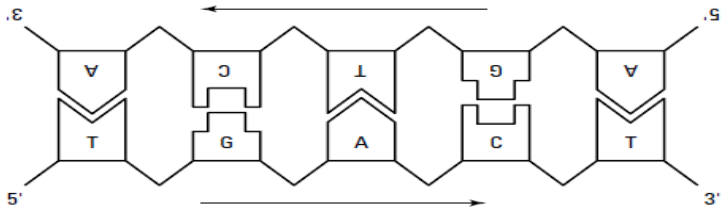
کد آیوپاک نوکلئوتیدهای DNA

جدول زیر کدهای یک حرفی (کد آیوپاک) که برای نشان دادن توالی DNA استفاده می شوند را فهرست کرده است . آیوپاک همچنین برای مواردی که به دو یا سه نوکلئوتید مشکوک هستیم کدهای خاصی تعریف کرده است. در این جدول فقط کدهایی که بیشترین استفاده را دارند، ذکر شده اند.

1-Letter Code	*Nucleotide Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
T	Thymine	Pyrimidine
N	Any nucleotide (any base)	(n/a)
R	A or G	Purine
Y	C or T	Pyrimidine
--	-----	None (gap)

جدول ۲-۱: معمول ترین کدهای مورد استفاده برای نوشتن توالی های DNA

وقتی گفته می شود دو رشته DNA مکمل همدیگر هستند یعنی تیمین (T) همواره با آدنین (A) و گوانین (G) همراه با سیتوزین (C) جفت می شود. این جفت نوکلئوتیدها (G-C, A-T) اگرچه با پیوند شیمیایی به هم متصل نشده اند اما رابطه مکملی بسیاراختصاصی دارند. به دلیل وجود این رابطه مکملی وقتی توالی یک رشته DNA مشخص باشد توالی رشته مقابل را می توان به راحتی حدس زد .



شکل ۶-۱: دو رشته مکمل یک مولکول DNA

این ویژگی شگفت انگیز- و نه ساختمان مارپیچی شیک- همه چیز را در رابطه با توالی های DNA شرح می دهد. به عنوان مثال زمانیکه یک ارگانیسم های زنده تولید مثل می کنند هر کدام از ژن های آنها باید دو برابرشوند . اما فرآیند دوبرابر شدن ژن ها مشابه فرآیندی نیست که در ماشین زیراکس انجام می شود یعنی یک رونوشت دقیق از ژن ها ساخته نمی شود، بلکه دو رشته DNA از هم جدا شده و با استفاده از رابطه مکملی بین نوکلئوتیدها در مقابل هر تک رشته DNA یک رشته مکمل ساخته می شود. دو رشته بودن DNA نشان دادن توالی DNA را روی کاغذ مشکل می کند. مطابق قرارداد، نوکلئوتیدها باید از انتهای 5′ به 3′ خوانده شوند اما توالی نوکلئوتیدی کدام یک از دو رشته DNA را باید خواند؟ جواب این است که توالی نوکلئوتید ها از هر کدام از دو رشته DNA که خوانده شود درست است. بنابراین هر قطعه از DNA می تواند با دو توالی (کاملاً متفاوت) تعریف شود. مطلب پیچیده ای نیست : فقط کافی است هنگام کارکردن با توالی های DNA این موضوع را به خاطر داشته باشید.

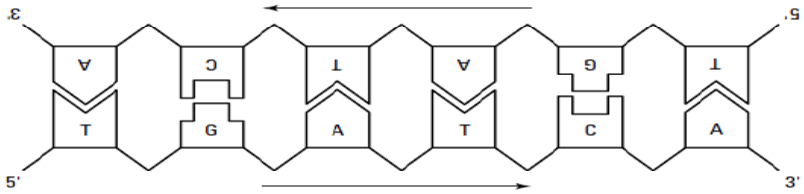
خوشبختانه اکثر برنامه های موتورهای جستجو بانک های اطلاعاتی مثل BLAST ، این ویژگی را می دانند و هر دو رشته مربوط به یک مولکول DNA را در نتایج جستجوی خود نشان می دهند. اما بعضی برنامه ها به خودشان زحمت نمی دهند و فقط توالی را که شما به برنامه وارد کرده اید را بررسی می کنند . زمانی که دو رشته DNA برای شما اهمیت دارند اطمینان حاصل کنید که یک آنالیز کامل از توالی مورد نظر شما صورت گرفته باشد(در فصل ۷-۵-۳ در مورد جزئیات این موضوع بیشتر بحث خواهد شد).

توالی هایی از DNA که از دو طرف یکسان خوانده می شوند

توالی های مکمل معکوس معمولاً برای افرادی که در آنالیز توالی DNA تازه کار هستند گیج کننده است. اما زمانی می رسد که شما هم با دید دو توالی زیر فوراً تشخیص دهید که مربوط به دو رشته متقابل یک مولکول DNA هستند. یک ویژگی فوق العاده DNA این است که در مناطقی از این مولکول توالی نوکلئوتیدها از روی هر یک رشته ها که خوانده شود (براساس قرارداد) یکسان خواهد بود . شکل ۷-۱ به فهم این مطلب کمک می کند.

چنین توالی هاییPalindromes نامیده می شوند. این توالی ها (مثل کلمات یا عباراتی که از دو طرف به یک صورت خوانده می شوند. مانند :"Madam, I'm Adam" یاPanama": "Aman , aplan azanal") فقط یک چیز عجیب

و غریب نیستند بلکه در زیست شناسی وظایف مهمی را بر عهده دارند. به عنوان مثال بعضی از این توالی ها محل اثر آنزیم های برش دهنده مولکول DNA (که آنزیم های محدود کننده[1] نام دارند) هستند. بعضی دیگر محل های اتصال پروتئین های تنظیم کننده هستند که باعث خاموش و روشن شدن یک ژن می شوند. این توالی ها همچنین نقش مهمی در تعیین شکل سه بعدی مولکول DNA دارند (البته نه فقط DNA؛ در بخش بعد در مورد این گونه توالی ها در RNA هم بحث خواهیم کرد.). جستجوی دقیق یا تقریبی این توالی ها در مولکول DNA می تواند یک تمرین عالی در زمینه بیوانفورماتیک باشد.



شکل۷-۱: دو رشته مکمل به صورت مشابهی خوانده می شوند

آنالیز توالی های RNA

DNA (دئوکسی ریبونوکلئیک اسید) مهمترین عضو خانواده ماکرو مولکول های نوکلئیک اسیدی است. وظیفه منحصر به فرد DNA در یک ارگانیسم حفاظت دائمی از اطلاعات ژنتیکی است. بنابراین مولکول DNA بسیار پایدار و مقاوم است و به خوبی درون هسته هر سلول محافظت می شود . ریبونوکلئیک اسید (RNA) عضو بسیار فعال تر خانواده نوکلئیک اسید است. این مولکول ساخته شده به محض این که کپی ژن های سلول را در اختیار کارخانه پروتئین سازی سلول قرارداد، تخریب می شود. در بیوانفورماتیک فقط دو تفاوت عمده DNA و RNA وجود دارد :

- ✓ DNA و RNA در یک نوکلئوتید با هم تفاوت دارند.
- ✓ RNA بر خلاف DNA تک رشته ای است.

کدهای یک حرفی آیوپاک برای نشان دادن توالی های RNA در جدول ۳-۱ فهرست شده اند.

1-Letter Code	*Nucleotide Base Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
U	Uracil	Pyrimidine
N	Any nucleotide	Purine or Pyrimidine

[1] restriction enzymes

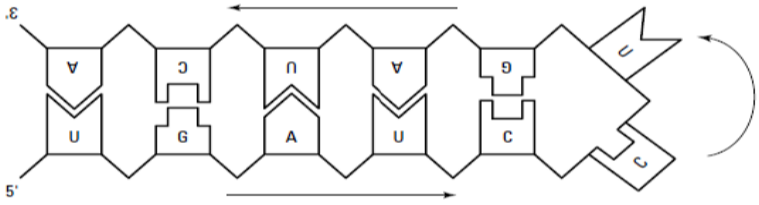
1-Letter Code	Nucleotide Base Name	Category
R	A or G	Purine
Y	C or U	Pyrimidine
--	-------	None (gap)

جدول ۳-۱: معمول ترین کدهای مورد استفاده برای نوشتن توالی RNA

بعضی برنامه ها به صورت خودکار U را با T جابجا می کنند اما بسیاری از برنامه ها حتی نمی توانند دو دسته از نوکلئوتیدها را از هم تمییز دهند . بنابراین اگر در یک بانک اطلاعاتی مشاهده کردید که در توالی یک RNA به جای U، T نوشته شده است تعجب نکنید . در حقیقت توالی های RNA با حروف الفبای DNA نوشته می شوند. به همین دلیل محققین به کار کردن با توالی های " ژن های " RNA (به زبان DNA) عادت کرده اند.

ساختار RNA : بازی کردن با رشته های چسبنده

اگرچه مولکول های RNA تک رشته ای هستند اما با این حال تمایل خود را به جفت شدن با رشته مکمل را هنوز حفظ کرده اند . حال مولکول تک رشته RNA را یک تکه نوار چسبدر نظر بگیرید که در محیط شناور است. واضح است در مدت کوتاهی بخش های چسبنده این نوار چسب به هم می چسبند و نهایتاً به شکل یک توده در می آیند. دقیقاً همین اتفاق برای یک مولکول تک رشته RNA هم می افتد. شکل ۸-۱ رفتار انتهاهای چسبنده را نشان می دهد.



شکل ۸-۱ : چگونه یک مولکول RNA دو رشته ای می شود.

اکنون اهمیت وجود رابطه مکملی بین رشته نوکلئیک اسید بیشتر روشن می شود. نواحی مختلف مولکول RNA تک رشته ای با هم جفت می شوند تا به فرم پایدار مارپیچ های مضاعف درآیند. اگرچه این مارپیچ های مضاعف کاملاً شبیه مارپیچ های مضاعف مولکول DNA هستند اما مسلماً از نظم کمتری برخوردار هستند. پس ازسنتز یک مولکول RNA، نوکلئوتید های آن تا حد ممکن با هم جفت شده یک توده متراکم و پرپیچ و تاب به وجود می آورند و طی همه این مراحل رشته RNA نه تنها انعطاف پذیری بلکه شکل هندسی خاص خود را هم حفظ می کند. اشکال شبیه سنجاق سر که در شکل ۸-۱ نشان داده شده اند عامل اصلی در تشکیل ساختمان دوم RNA هستند. این اشکال هر کدام از یک لوپ (در شکل ۸-۱ ، C, U که به هم جفت نشده اند) و یک شاخه (نوکلئوتیدهایی که با هم جفت شده اند) تشکیل شده اند.

برای سرگرمی این سؤال را بررسی کنید که آیا یک توالی Palindrome در RNA می تواند بدون تشکیل لوپ یک سنجاق سر کامل را به وجود آورد؟

وقتی نوکلئوتیدهای یک مولکول RNA تا حد ممکن با هم جفت شوند، بخش های مختلف RNA روی هم تا می خورند و در نهایت یک ساختار سه بعدی اختصاصی برای آن مولکول حاصل می شود. توالی نوکلئوتیدی هر مولکول RNA در نهایت ساختار سه بعدی نهایی آن را تعیین می کند. نقش بیولوژیکی مولکول های RNA از ساختار سه بعدی آنها و یا از مکمل بودن آنها با ژن های خاص ناشی می شود. پیش بینی ساختار نهایی RNA از روی توالی نوکلئوتیدی آن یک چالش بزرگ است کهدر گذشته باعث بسیاری از تحولات در بیوانفورماتیک شده است. اخیراً کشف این مسئله که مولکول های کوچک RNA می توانند فعالیت تعدادی ژن را متوقف کنند، دوباره باعث ایجاد علاقه به این توالی چسبنده شده است (اگر به بیوانفورماتیک مولکول RNA علاقه مند شده اید می توانید مستقیماً به قصل ۱۲ مراجعه کنید).

اطلاعات بیشتر در مورد نامگذاری اسید های نوکلئیک

اگر مشاهده می کنید کتاب ها، دوره ها یا مقالات مختلف، از کلمات یا علامات اختصاری متفاوتی برای نشان دادن نوکلئیک اسیدها استفاده می کنند، اصلاً نترسید! شما ممکن است با کلماتی مثل "باز" ، " جفت باز " ، "نوکلئوتید" یا " نوکلئوزید" مواجه شوید اما توجه کنید که این ترکیبات شیمیایی تفاوت خیلی کمی با هم دارند و این تفاوت ها الان برای ما مهم نیستند. در این کتاب تا به حال از کلمه " نوکلئوتید" استفاده کرده ایم که علامت اختصاری آن nt است. مثلاً برای اشاره به یک توالی RNA یا DND به طول ۴۰۰ نوکلئوتید می نویسیم: "a 400-nt –long sequence". در این روش نامگذاری، طول یک مولکول RNA یا DNA بر حسب تعداد موقعیت های موجود برای نوکلئوتیدها بیان می شود . مثلاً برای اشاره به توالی شکل ۴-۱ می نویسیم:

"a 5–nt–long sequence"

توجه کنید که در نامگذاری یک توالی تعداد موقعیت های موجود برای نوکلئوتیدها بیان می شوند نه تعداد نوکلئوتیدها. وقتی می گوییم یک مولکول DNA ، " 400–nt–long " است یعنی ۴۰۰ موقعیت برای نوکلئوتید دارد اما تعداد واقعی نوکلئوتیدهای آن دو برابر این رقم است یعنی ۸۰۰، چون هر موقعیت در این مولکول حاوی یک نوکلئوتید است. برای این که این ابهام برطرف شود ، اندازه توالی DNA اغلب به صورت جفت باز (bp) بیان می شوند. بدین ترتیب طول توالی DNA شکل ۵-۱ ، 5 bp خواهد بود. واحدهای بزرگتر مثل kb (1000 bp) یا Mb (Mega-bp) استفاده می شوند.

نواحی کد کننده پروتئین: شبیه سازی کار با توالی های آمینواسیدی

تنها درصد کمی از صدهزار توالی پروتئینی که در حال حاضر در بانک های اطلاعاتی وجود دارند مستقیماً جداسازی و تعیین توالی شده اند. چون همان طور که قبلاً هم گفته شد تعیین توالی پروتئین بسیار مشکل تر از DNA است. ولی همه پروتئین هایی که یک ارگانیسم (خواه باکتری باشد یا انسان) می تواند سنتز کند توسط DNA موجود در ژنوم آن ارگانیسم کد می شوند. بنابراین

زیست شناسان از یک راه میانبر برای توالی یابی پروتئین ها از روی مولکول DNA کد کننده آنها! استفاده می کنند. بدین ترتیب ما می توانیم ادعا کنیم که توالی آمینواسیدی یک پروتئین را بدون این که حتی آن را جداسازی کنیم، می دانیم.

تبدیل DNA به پروتئین : کد ژنتیکی

با دانستن کدهای ژنتیکی، می توان توالی نوکلئوتیدیDNA را به توالی آمینو اسیدی پروتئین ترجمه کرد. سلول هم به همین ترتیب یک توالی پروتئینی را از روی مولکول DNA می سازد. تنها تفاوت کار ما با کار سلول این است که ما به جای مواد شیمیایی واقعی از حروف اختصاری استفاده می کنیم. کدهای ژنتیکی عمومی هستند و در همه موجودات یکسانند و در طبیعت همین کدهای ژنتیکی هستند که یک ارتباط منحصر به فرد بین توالی DNA و توالی پروتئینی ایجاد می کنند (البته با چند استثنا وگرنه حیات موجودات خیلی ساده می شد!). این مطلب یکی از بزرگترین دستاوردهای زیست شناسان در دهه ۱۹۶۰ بود. کدهای ژنتیکی شکل ۹-۱ به صورت یک جدول آمده اند:

نحوه استفاده از جدول مذکور : از یک روی DNA شروع کرده و دسته های سه تایی از نوکلئوتیدها (که کدون نامیده می شوند) درست می کنیم. سپس آمینواسید مربوط به هرکدون را از جدول پیدا می کنیم. مثلاً برای توالی DNA (یا DNA پیام آور) زیر بدین ترتیب عمل می کنیم:

۱. خواندن توالی DNA

ATGGAAGTATTTAAAGCGCCAVVTATTGGGATATAAG

۲. تفکیک دسته های سه تایی (کدون ها)

ATG GAA GTA TTT AAA GCG CCA CCT ATT GGG ATA TAA G...

۳. ترجمه هر کدون به آمینواسید مربوطه

M E V F k A P Q I G I S T O P

اگر جهت توالی DNA از 5 به 3′ باشد توالی پروتئین حاصل نیز از انتهای N به Cخواهد بود . مزیت این روش این است که دیگر لازم نیست نگران جهت توالی پروتئین باشیم. بنابراین اگر نقطه ابتدایی که ناحیه کد کننده پروتئین روی مولکول DNA را بدانید کامپیوتر شما می تواند مثل یک سلول زنده توالی آمینواسیدی مربوطه را ایجاد کند. اکثر" توالی های پروتئین " موجود در بانک های اطلاعاتی با همین تمرین ساده " ترجمه توسط کامپیوتر " به وجود آمده اند . بسیاری از برنامه های توالی یابی از این موضوع استفاده کرده و " ترجمه در پرواز " [1] را ارائه کردند ، بنابراین با یک کلیک می توانید توالی DNA را به توالی مجازی پروتئین تبدیل کنید.

[1] On-the-fly translation

در مورد کد شدن توالی DNA به پروتئین بیشتر بدانیم

اگر به مثالی که در بخش قبل ذکر شد دقت کنید متوجه خواهید شد که توالی پروتئینی حاصل کاملا به این موضوع بستگی دارد که چگونه توالی DNA به دسته های نوکلئوتیدی سه تایی دسته بندی شود . مثلاً اگر در مثال مذکور نقطه شروع ما نوکلئوتید دوم باشد:

1- A**TGG**AAG**TAT**TTA**AAG**CGC**CAC**CTA**TTG**GGA**TAT**AAG

2- A **TGG** AAG **TAT** TTA AAG CGC **CAC** CTA **TTG** GGA **TAT** AAG

3- W K Y L K R H L L G Y K

به همین ترتیب اگر از نوکلئوتید سوم شروع کنیم نتیجه کاملاً متفاوت بود.

	T	C	A	G
T	TTT Phe (F) TTC Phe (F) TTA Leu (L) TTG Leu (L)	TCT Ser (S) TCC Ser (S) TCA Ser (S) TCG Ser (S)	TAT Tyr (Y) TAC Tyr (Y) TAA Stop TAG Stop	TGT Cys (C) TGC Cys (C) TGA Stop TGG Trp (W)
C	CTT Leu (L) CTC Leu (L) CTA Leu (L) CTG Leu (L)	CCT Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAT His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGT Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)
A	ATT Ile (I) ATC Ile (I) ATA Ile (I) ATG Met (M)	ACT Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAT Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGT Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)
G	GTT Val (V) GTC Val (V) GTA Val (V) GTG Val (V)	GCT Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAT Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGT Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)

شکل ۹-۱: کدهای ژنتیکی عمومی

چون کدهای ژنتیکی سه تایی هستند از نظر تئوری یک قطعه از یک رشته DNA به سه شکل می تواند خوانده شود(قالب های خواندن[1]). و چون در فرآیند ترجمه از هر دو رشته DNA می تواند استفاده شود ۶ حالت کلی برای ترجمه یک توالی DNA به یک توالی پروتئین ممکن خواهد بود. از این ۶ حالت ممکن به جز در موارد بسیار نادری که استثنا وجود دارد (ویروس های

[1] reading frames

نامتعارف[1])، فقط یکی از حالت ها برای یک منطقه کدکننده پروتئین در مولکول DNA اتفاق می افتد. قطعه ای از یک رشته DNA که فاقد کدون خاتمه (TAA, TGA, TAG) باشد، قالب باز خواندن[2] (ORF) نامیده می شود.

ارگانیسم های متکامل تر قطعات بزرگی از DNA در ژنوم خود دارند که اصلاً به پروتئین ترجمه نمی شوند. این قطعات که اینترون[3] نام دارند باعث پیچیده تر شدن فرآیند ترجمه می شوند. به همین دلیل بخش قابل توجهی از بیوانفورماتیک به یافتن روش هایی برای تعیین محل های کد کننده پروتئین روی مولکول DNA اختصاص داده شده است. بدین ترتیب می توان نقاط شروع و پایان ژن ها یا مناطق غیر کد کننده (اینترون ها) را به دقت تعیین کرد.

سرفصل های مرتبط با بیوانفورماتیک **DNA** و **RNA** در این کتاب

- ✓ بازیابی توالی DNA از بانک های اطلاعاتی (فصل ۳-۲)
- ✓ تعیین نوکلئوتیدهای متشکله DNA (فصل۵)
- ✓ مشخص کردن نواحی محدود کننده[4] (فصل ۵)
- ✓ طراحی پرایمر برای PCR[5] (فصل۵)
- ✓ مشخص کردن نواحی ORF (فصل۵)
- ✓ پیشگویی عوامل اصلی تعیین کننده ساختمان دوم RNA, DNA (فصل ۱۲)
- ✓ یافتن توالی های تکراری (فصل۵)
- ✓ یافتن بهترین تطابق بین دو یا تعداد بیشتری توالی RNA, DNA (فصل۷، ۸ و ۹)
- ✓ یافتن نواحی پلی مورف در ژن ها (پلی مورفیسم های تک نوکلئوتیدی[6] یا SNPs) (فصل۳)
- ✓ قطعاتی که باعث تجمع توالی می شوند.(فصل۵)

کارکردن با ژنوم کامل

در سال ۱۹۷۷ برای اولین بار یک تکنیک کارا برای تعیین توالی DNA کشف شد و در سال۱۹۹۵ برای اولین بار ژنوم یک ارگانیسم (میکروب هموفیلوس آنفلوانزا) به طور کامل تعیین توالی شد. در سال های بین ۱۹۷۷و ۱۹۹۵ تکنیک های تعیین توالی DNA دائماً در حال پیشرفت بودند اما این تکنولوژی ها بیشتر در جهت استخراج ژن هاو به دست آوردن اطلاعات از آنها به کار

[1] exotic viruses
[2] open reading frames
[3] introns
[4] restriction stites
[5] Palymerase chain- reaction
[6] single nucleotide polymorphisms

برده شدند. در این دوره زیست شناسان چون کار کردن روی این ژن ها را از چند سال قبل شروع کرده بودند اغلب ژن های به طول چند هزار نوکلئوتید را تعیین توالی می کردند. اغلب ابزارهای موجود بیوانفورماتیک از آن زمان به وجود آمده اند. بعنوان مثال:

- ✓ همه برنامه هایی که توالی را با هم تطبیق می دهند[1].
- ✓ روش های فیلوژنیک و روش های دسته بندی
- ✓ ابزارهای نمایشگر مختلف که برای توالی های نسبتاً کوچک (مثلاً توالی های پروتئینی که طول آنها کمتر از چند هزار حرف است) مناسب هستند.

ژنومیک: بدست آوردن همزمان همه ژن ها یا ژنوم کامل

تعیین توالی کامل یک ژنوم برای اولین بار، باعث شد که روش تعیین توالی ژن های منفرد کنار گذاشته شود و نقطه آغازی برای ژنومیک، تعیین نقشه ژنتیک و تعیین نقشه فیزیکی و تعیین توالی کامل ژنوم بود. در نتیجه توالی های DNA که امروزه روی آنها کار می شودند بسیار بلندتر هستند، مثلاً ژنوم میکروب ها حدود 1,000,000 bp و ژنوم حیوانات و انسان تا چندین میلیارد جفت باز طول دارد. این انقلاب مستلزم طراحی ابزارها و بانک های اطلاعاتی جدید بیوانفورماتیک است که قادر باشند این حجم بزرگ اطلاعات را ذخیره ، بازرسی و آنالیز کرده به شکلی مطلوب نمایش دهند. در فصل های ۷-۵-۳ سؤالاتی که زیست شناسان در کارکردن در مقیاس ژنوم باآنها مواجه می شدند و ابزارهای بیوانفورماتیک مربوطه ارائه شده اند.

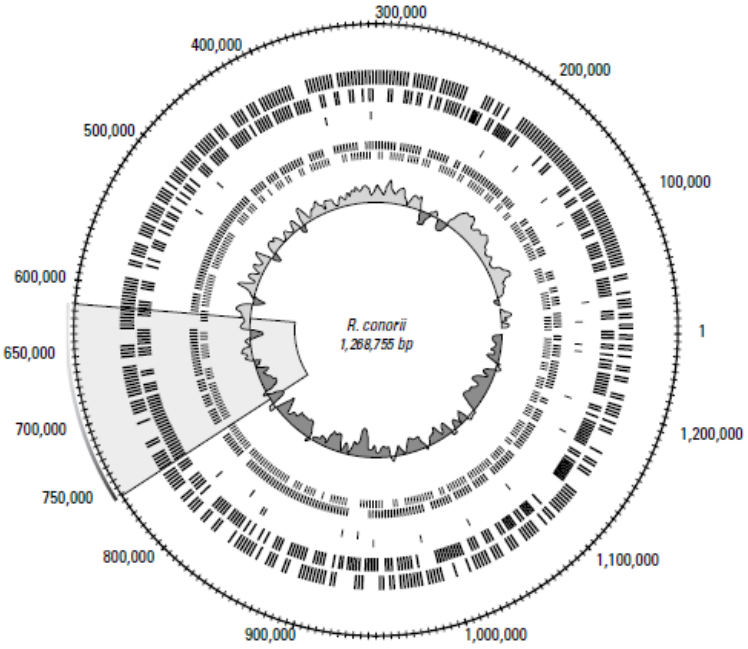
بر خلاف روزهای اولیه تعیین توالی ژن ها منفرد، امروزه توالی های DNA (و همزمان توالی های پروتئین مربوطه به آن توالی های DNA) اغلب بدون هیچگونه دانش قبلی از آنها تعیین می شوند یعنی ژن ها همزمان هم تعیین توالی و هم کشف می شوند. این تحول باعث پیدایش یک شاخه کاملاً جدید از بیوانفورماتیک شد که به تجزیه توالی های بزرگ DNA به اجزای تشکیل دهنده آنها (ژن ها ، واحدهای نسخه برداری ، نواحی کد کننده پروتئین ، عوامل تنظیمی و ...) اختصاص پیدا می کند. در مرحله بعد نقش بیولوژیکی این اجزاء کم و بیش پیش بینی می شود. در فصل ۴ این کتاب به بعضی از پیشرفته ترین تکنیک ها دراین زمینه اشاره می شود.

شکل ۱۰-۱ که مربوط به ژنوم کامل باکتری Rickettsia conorii است، این شاخه جدید بیوانفورماتیک رانشان می دهد این مولکول DNA حلقوی 1.3 میلیون جفت باز طول دارد. هر مستطیل کوچک در دودایره خارجی تر ، یک ناحیه کد کننده پروتئین است و تقریباً 1000 bp طول دارد. قبل از این که تعیین توالی ژنوم این باکتری انجام شود هیچکس نمی دانست چه ژن هایی (یا چه پروتئین هایی) در این باکتری وجوددارد. همه اطلاعاتی که ما اکنون در مورد این باکتری (و بسیاری دیگر از باکتری هایی که نسبتاً غیر قابل دسترسی هستند مثل باکتری هایی که در آب های با دمای $100^{\circ}C$ در نزدیکی دهانه آتشفشان ها) داریم، از آنالیز بیوانفورماتیک حاصل شده اند.

[1] sequence-alignment programs

سرفصل های مرتبط با بیوانفورماتیک ژنوم

- ✓ یافتن این که چه ژنوم هایی موجودند (فصل ۳)
- ✓ آنالیز توالی هایی که با ژنوم های خاصی مرتبط هستند(فصل ۷-۳)
- ✓ نشان دادن ژنوم ها (فصل ۳)
- ✓ تجزیه توالی یک ژنوم میکروبی یا تعیین نواحی ORF[1] (فصل ۵)
- ✓ تجزیه توالی یک ژنوم یوکارتی (GenScan) (فصل ۵)
- ✓ یافتن ژن های ارتولوگ[2] و پارالوگ[3] (فصل ۳)
- ✓ یافتن توالی های تکراری



شکل ۱۰-۱: نمایش یک ژنوم باکتریایی

[1] ORFing
[2] Orthologous
[3] Paralogous

فصل دوم

چگونه از بیوانفورماتیک استفاده کنیم؟

در این فصل می خوانیم :

- ✓ استفاده از PubMed برای بدست آوردن سریع اطلاعات در مورد موضوعات بیولوژیک
- ✓ بازیابی توالی DNA یا پروتئین
- ✓ انجام اولین مقایسه(تطبیق) بین توالی و بانک اطلاعاتی با استفاده از BLAST
- ✓ انجام اولین تطبیق بین چند توالی با استفاده از ClustalW

.....

" اگر چه من همیشه دوست ندارم آموزش داده شوم، اما همیشه دوست دارم یاد بگیرم."

— Sir Winston Churchill

در این فصل شما به اندازه کافی دانش بیوانفورماتیک برای انجام کارهای روزمره (اما بسیار مفید) خود خواهید آموخت. کارهایی که زمانی زیست شناسان با تجربه ماه ها پشت اینترنت وقت صرف انجام آنها می کردند. فقط با دانستن مطالبی این فصل شما می توانید ۴۰ درصد نیازهای حرفه ای خود را تأمین کنید. کافی است به اندازه کافی مطالیب که تاکنون گفته شد را آموخته باشید تا مطالب مفیدی را از این فصل یاد بگیرید. اگر شما سابقه کار در علوم کامپیوتر را دارید، ما در این فصل ابزارهایی که زیست شناسان نیاز دارند را به شما نشان خواهیم داد.

چگونه می توان به سرعت یک متخصص PubMed/ Medline شد؟

در نظر بگیرید شما توالی ژنی را که سال ها روی آن کار می کردید را به دست آورده اید وآن را به یک متخصص بیوانفورماتیک بد اخلاق داده اید تا آن را آنالیز کند (البته این اتفاق فقط قبل از خواندن این کتاب ممکن است براسی شما بیفتد.) تشخیص متخصصین اینگونه است "توالی شما تقریباً شبیه dUTPase است."

گرفتن اطلاعات در مورد یک پروتئین با استفاده از نام آن

بهترین حالت این است که ژن شما شبیه به چیزی باشد که در این صورت می تواند نقطه شروع یک داستان باشد. اما تصور کنید که شما در طول زندگی تان حتی اسم dUTPase را هم نشنیده باشید؛ حالا چه کار خواهید کرد؟ چه طور می توانید در مورد این کلمه ظاهراً مشکوک و عجیب و غریب اطلاعات بیشتری به دست آورید؟ چگونه می توانید به سرعت تشخیص دهید که آیا این یافته جدید با اطلاعاتی که شما قبلاً از آن ژن داشتید هم خوانی دارد یا نه؟

جستجو کردن در PubMed

جواب سؤال های بالا این است که روی صندلی خود بنشینید و به سایت PubMed وارد شوید. برای وارد شدن به PubMed بدین ترتیب عمل کنید:

۱. با استفاده از جستجوگر اینترنت، وارد آدرس اینترنتی زیر شوید.

http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed

اکنون صفحه کامپیوتر شما شبیه شکل ۱-۲ خواهد بود. در بالای صفحه یک پنجره وجود دارد که از قبل لغت PubMed در آن نوشته شده و یک پنجره جستجو در کنار آن که شما باید کلمات کلیدی مربوط به سؤال خود را آن جا تایپ کنید.

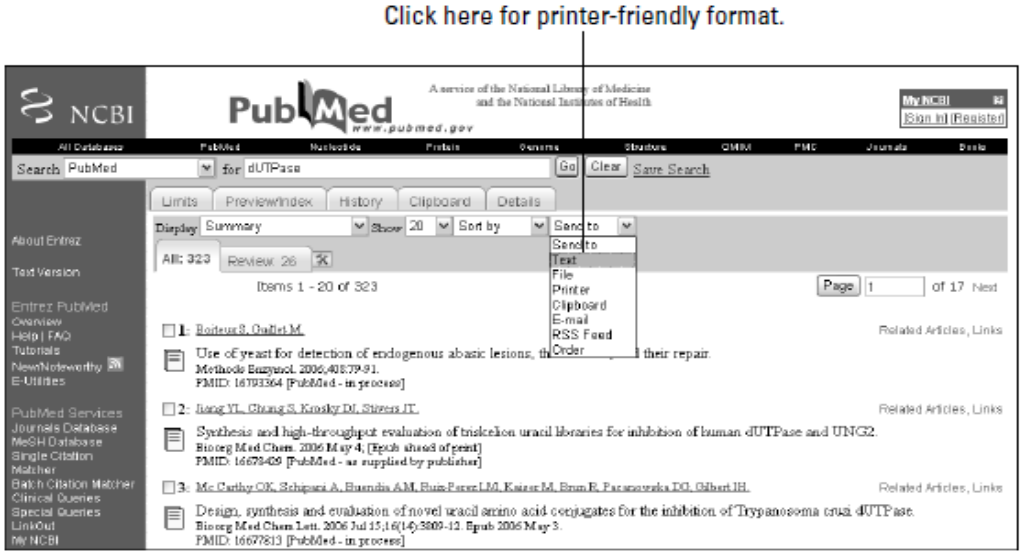
۲. dUTPase را در پنجره for تایپ کنید و سپس روی دکمه Go کلیک کنید. چند لحظه بعد لیستی از نتایج جستجو شبیه به شکل ۲-۲ ظاهر می شود. در مورد مثال dUTPase در نتیجه جستجو بیش از ۲۰۰ مرجع خواهیم داشت. مسلم است که پیدا کردن راه حل مشکل ما در این ۲۰۰ مرجع کار مشکلی است و اما اشتباه نکنید، لازم نیست یک پرینت از این لیست را در دست بگیرید و در کتابخانه دنبال آنها بگردید!

شکل ۱-۲: پنجره اولیه جستجوی PubMed

۳. در هر مورد از نتایج بدست آمده روی نام مؤلف هاکلیک کنید. نام های مؤلف ها به رنگ آبی هستند و خطی زیر آنها کشیده شده است (این دو علامت نشان دهنده متن هایی هستند که با کلیک روی آنها می توان به متن دیگری که به آن پیوند شده است دسترسی پیدا کرد.)

با کلیک کردن روی این بخش صفحه ای باز می شود که حاوی خلاصه ای از مقاله ای است که شما انتخاب کرده اید . اگر یک زمینه علمی قابل قبول در زیست شناسی و کمی هم استعداد داشته باشید خواندن تعداد کمی از این خلاصه مقالات کافی است تا یک ذهنیت کلی در مورد dUTPase پیدا کنید.

۴. با کلیک روی گزینه file در بالای صفحه و سپس کلیک روی Save as مطالبی که برای شما مفیدند را ذخیره کنید. یا اینکه می توانید این صفحه نمایش را به شکل ساده تری که برای پرینتر مناسب تر (printer friendly) باشد، تبدیل نمایید. برای این کار از منوی باز شونده از پایین (Send to (Drop-down، Text را انتخاب نمایید و سپس در File مرورگر، Save as را انتخاب کنید.



شکل ۲-۲: نتیجه اولیه جستجوی معمولی کلید واژه dUTPase در PubMed

چگونه چند نتیجه جستجوی دلخواه را ذخیره کنیم؟

زمانی که نتایج جستجو زیاد بودند بهترین کاری که می توان کرد این است که به چند خلاصه مقاله مراجعه کنید و هر کدام که برای شما مفید بودند را با کلیک کردن در مربع کنار آن انتخاب کنید. برای رفتن به صفحه بعدی نتایج کافی است روی کلمه next در بالای سمت راست صفحه کلیک کنید. زمانی که متوجه شدید که به اندازه کافی مقاله انتخاب کرده اید ، مراحل زیر را انجام دهید:

۱. از پنجره Display گزینه Abstract را انتخاب کنید (با کلیک روی مربع کنار پنجره یک لیست برای شما باز می شود که می توانید Abstract را از آن لیست انتخاب کنید.) در این حالت خلاصه همه مقالاتی که شما انتخاب کرده اید ظاهر می شود.
۲. برای پرینت کردن این خلاصه مقالات می توانید از file گزینه print را انتخاب کنید.
۳. اگر می خواهید خلاصه مقالات را به یک فرم مناسب تر پرینت کنید در پنجره ای که در حالت عادی درون آن send to نوشته شده، گزینه text را انتخاب کنید (شکل ۲-۲) . در این صورت اطلاعات شما دیگر به فرم تحت وب نخواهد بود. سپس با انتخاب گزینه print از پنجره file می توانید پرینت بگیرید.
۴. اگر می خواهید اطلاعات خود را ذخیره کنید کافی است در منوی file گزینه save as را انتخاب کنید.

اگر چه در پنجره sent to هم گزینه file وجود دارد اما چون گاهی اوقات این گزینه با تنظیمات ایمنی کامپیوتر تداخل پیدا می کند بهتر است از منوی استاندارد file که در بالای سمت چپ صفحه قرار دارد استفاده کنید.

جستجو در PubMed با استفاده از نام مؤلف

اگر چه شما می توانید به راحتی با استفاده از کلمات کلیدی موضوع مورد نظرتان (مثلاً نام پروتئین) PubMed را جستجو کنید ، اما گزینه های دیگری هم برای جستجو کردن وجود دارد. مثلاً گاهی ممکن است همکار یا استاد راهنامی شما به شما بگوید: من مقاله ای به وسیله فلانی و فلانی نوشته شده و مدت زیادی از چاپ آن نمی گذرد را خوانده ام و فکر می کنم در مورد dUTPase باشد، "بد نیست آن را هم بخوانی" از منظردنیای علمی قدیم برای پیدا کردن چنین مقاله ای باید فهرست ها و قفسه های کتابخانه را یکی یکی بگردیم که کاری تقریباً محال به نظر می رسد. اما در اینترنت PubMed با شماست. فرض کنید شما فقط نام یکی از مؤلف های یک مقاله، مثلاً Abergel، را به یاد دارید. کافی است مراحل زیر را انجام دهید:

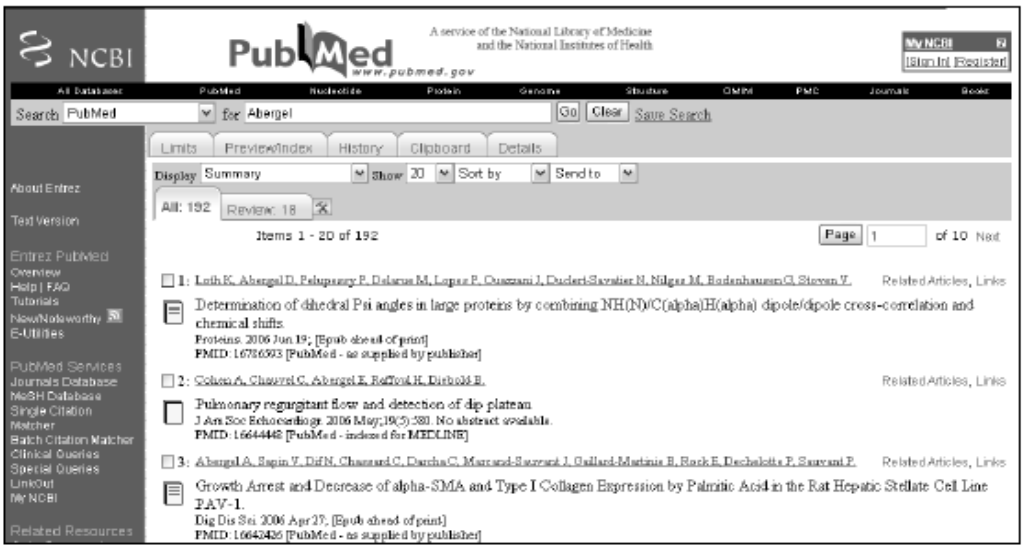
۱. آدرس اینترنتی www.ncbi.nlm.nih.gov/entrez/ را جستجو کنید .
۲. در پنجره for لغت Abergel را تایپ کنید. سپس روی دکمه Go کلیک کنید. نتیجه جستجوی شماچیزی شبیه شکل ۳-۲ خواهد بود.

حالا ما یک لیست ازمقالاتی داریم که Abergel یکی از مؤلف های آن است، به همان ترتیبی که قبلاً گفته شد می توانید از بین آنها مقالاتی را که برای شما مفید است انتخاب کنید و به خلاصه آن مقالات دسترسی پیدا کنید. بدترین

حالت ممکن می تواندباین باشد که هیچکدام از این مقالات ارتباطی با dUTPase نداشته باشند. برای حل این مشکل می توانید نام پروتئین و مشخصات مؤلف را همزمان جستجو کنید.

۳. در پنجره for لغات dUTPase و Abergel را با یک فاصله بین آنها تایپ کنید سپس روی دکمه Go کلیک کنید.

در این حالت PubMed این گونه فرض می کندکه شما می خواهید هر دو لغت dUTPase و Abergel را جستجو کنید. همچنین PubMed فرض می کند که شما می خواهید این لغات در عنوان یا خلاصه مقالات جستجو شوند. شکل ۴-۲ نتیجه این جستجوی ترکیبی را نشان میدهد.

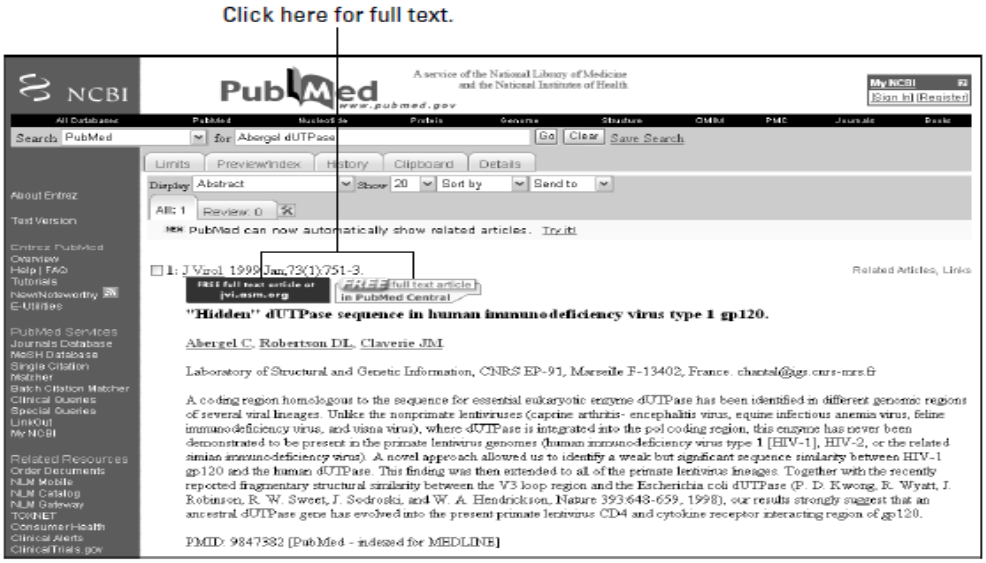


شکل ۳-۲ : نتایج جستجوی معمولی کلید واژه Abergel در Pub Med

چون فقط یک مقاله مرتبط با کلماتی که شما جستجو کرده اید وجود دارد PubMed خود به خود از حالت "Summary" به حالت "Abstract" در می آید و شما می توانید خلاصه تنها مقاله مرتبط با موضوع خود را بخوانید. این همان مقاله ای است که همکار یا استاد راهنمای شما در مورد آن صحبت می کرد. تبریک می گوییم ! شما الان اولین جستجوی موضوعی[1] خود را در PubMed انجام دادید.

زمانی که زمان جستجو بیش از یک مقاله بود به سادگی می توانید گزینه Abstract را در پنجره Display انتخاب کنید و خلاصه مقالات را مشاهده کنید.

[1] Bibliographical search



شکل ۴-۲: نتیجه جستجوی ترکیبی لغات dUTPase و Abergel در Pub Med

اما صبر کنید! انگار چیزهای دیگری هم در نتیجه جستجو وجود دارد . در حقیقت امروزه روز شانس شما بوده چون مقاله شما از ژورنالی است که امکان دسترسی رایگان به full text مقاله را در اختیار شما قرار داده است (در مستطیل بزرگ رنگی در بالای عنوان مقاله به این موضوع اشاره شده است.)

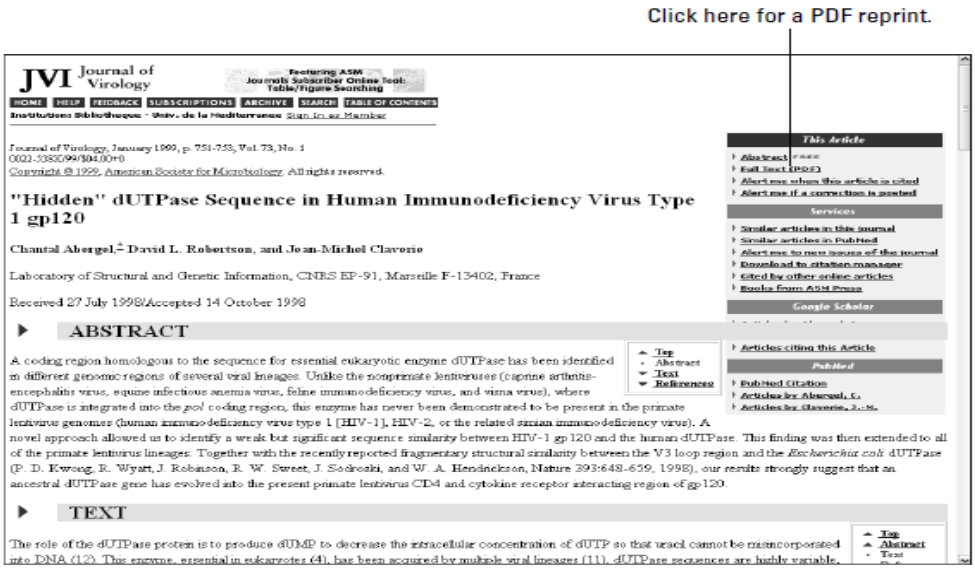
٤. روی مستطیل آبی رنگ کلیک کنید.

بااین کارمستقیماً وارد سایت ناشر این مقاله می شوید. در این سایت شما می توانید نسخه تحت وب مقاله خود را بخوانید یا یک پرینت از آن در قالب PDF بگیرید. (شکل۵-۲)

امروزه تعدا فزاینده ای از ژورنال های علمی از جمله بعضی ژورنال هایی که در زمینه کاری خودشان برجسته هستند full text مقالات را به صورت رایگان در اختیار قرار می دهند. اما به هر حال در اکثر ژورنال ها هنوز برای دسترسی به full text یک مقاله باید با پرداخت هزینه شخصاً مشترک آن ژورنال شد یا از طریق دانشگاهی که در آن ژورنال اشتراک دارد به آن مقاله دسترسی پیدا کرد .در بعضی ژورنال ها (مثلاً ژورنالی که در شکل ۵-۲ نشان داده شده) ۶ ماه پس از چاپ شدن مقاله، دسترسی به full text آن رایگان می شود. ضمناً اگر روی مستطیل دیگری که کنار مستطیل آبی است کلیک کنید به همان مقاله دسترسی پیدا می کنید . اما این بار از طریق انبار مرکزی PubMed[1].

بنابر این با وجود PubMed و ناشرهایی که با آن همکاری می کنند، اگر خوش شانس باشید می توانید بدون این که از صندلی خود بلند شوید، جستجوی موضوعی را در زمان کوتاهی انجام دهید و هر مطلبی که در مورد موضوع مورد نظرتان نیاز دارید، بدست آورید.

[1] PubMed Central



شکل ۵-۲: ژورنال Virology که امکان دسترسی رایگان به full text مقالات در آن وجود دارد.

به این نکته توجه داشته باشید که لغاتی که در پنجره for نوشته می شوند چه با حروف بزرگ باشند و چه کوچک تفاوت نمی کند؛ بنابراین جستجوی DUTPase ، dUTPase یا dutpase نتیجه یکسانی خواهد داشت. همچنین با لغات Abergel یا aBERGEI نتایج یکسانی حاصل می شود.) اما توجه داشته باشد که این موضوع در مورد همه بانک های اطلاعاتی و برنامه های جستجویی که در بیوانفورماتیک استفاده می شوند صادق نیست. (Case sensitivity)

جستجو در PubMed با استفاده از "field" ها

PubMed علاوه بر جستجو با استفاده از نام مؤلف یا موضوع مورد نظر، راه های مختلفی برای اختصاصی تر کردن جستجو ارائه می کند. در مدلاین تعداد زیادی ورودی[1] وجود دارد که PubMed نتایج جستجو را از آنها جمع آوری می کند. برای این که بتوانیم از راه های مختلفی که PubMed برای اختصاصی کردن جستجو دارد، استفاده کنیم، باید در مورد ساختار داخلی این ورودی ها اطلاعات بیشتری داشته باشیم. برای این که یک دید کلی از ساختارهای داخلی پیدا کنید، به شکل ۴-۲ مراجعه نمایید.

در قسمت قبل دیدیم که زمانی که دو لغت dUTPase و Abergel را به صورت ترکیبی جستجو کنیم PubMed فقط یک مقاله برای ما پیدا می کند. برای ادامه کار می توانید به صورت زیر عمل کنید:

۱. روی فلش کوچک کنار پنجره Display کلیک کنید تا منوی مربوطه باز شود. گزینه های موجود در این منو فرم های مختلف نمایش نتایج حاصل از جستجو یا ضمیمه های این نتایج هستند.
۲. گزینه MEDLINE را انتخاب کنید (شکل ۶-۲). صفحه مدلاین باز می شود (شکل ۷-۲).

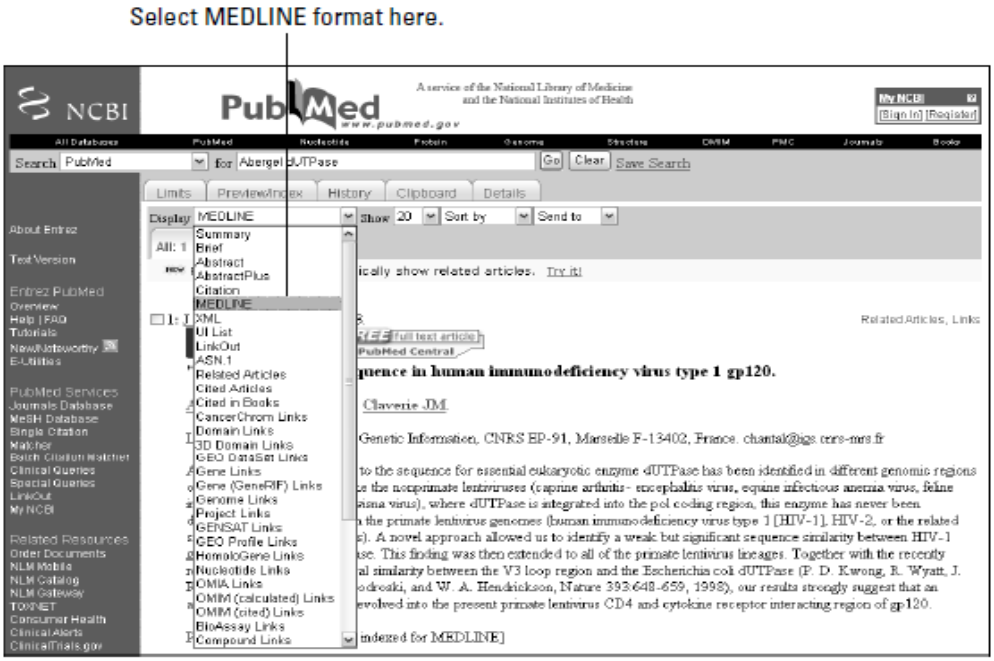
در شکل ۷-۲ می توانید ساختار داخلی پرونده یک مقاله که از بانک اطلاعاتی استخراج شده است را مشاهده کنید. در این صفحه اطلاعات در قسمت های مجزایی قرا گرفته اند که هر کدام یک "field " نامیده می شوند و با حروف اختصاری نمایش

[1] Entry

داده می شوند(مثلاً TT یعنی عنوان، AB یعنی خلاصه مقاله، AD یعنی آدرس آزمایشگاه، AU یعنی مؤلف ها، SO یعنی نام ژورنال و ...). این ساختار برای هر کدام از اطلاعاتی که از مدلاین استخراج می شوند وجود دارد.

استفاده درست از "field" ها

متأسفانه ، جستجوهای PubMed همیشه مانند جستجوی dUTPase آسان و نتیجه بخش نیستند. اگر کلمه ای که می خواهید جستجو کنید یک نام معروف باشد، (مثلاً smith یا Cohen) یا کلمه ای باشد که موارد استفاده زیادی داشته باشد مثلاً Down که می تواند در عنوان (Down syndrome)، خلاصه مقاله، نام مؤلف ها، و حتی در آدرس (955 down street) وجود داشته باشد، در این صورت تعداد نتایج جستجو [1] فوق العاده زیاد خواهد بود.



شکل ۶-۲: تبدیل فرم abstract به فرم MEDLINE

برای این که مشکل کمتر شود می توانیم هر کدام از کلماتی که می خواهید جستجو کنیم را در یک زمینه خاص محدود کنیم بدین ترتیب فقط مقالاتی که کمه مورد نظر در یک قسمت خاص از آنها وجود دارد جستجو می شوند. بدین منظور کافی است بعد از هر کلمه حروف اختصاری محلی که می خواهیم این کلمه در آن جستجو شود را در یک کروشه تایپ می کنیم. با این کار نتیجه جستجو کاملاً متفاوت خواهد بود. برای روشن شدن بیشتر مطلب کلمه Down را به سه حالت زیر جستجو کنید:

Down[AD], Down[TI], Down[AU]

برای سه حالت جستجوی بالا به ترتیب ۲۷۵ ، ۱۶۳۱۸ و ۱۲۱۳ نتیجه ی کاملاً غیر مرتبط به دست می آید.

[1] Hits لغت استاندارد برای نتایج جستجو است.

استفاده از "field " ها برای یافتن متخصصین یک زمینه خاص

جستجوی محدود به "field " ها زمانی که نام های مؤلف ها یا موضوع مورد نظر ما در "field " ها مختلفی یافت می شوند بسیار مفید است. اما این نوع جستجو کاربردهای زیاد دیگری هم دارد، به خصوص برای پیدا کردن متخصصین یک رشته خاص که در محل خاصی کار می کنند.

فرض کنید همین الان هواپیمای شما در شیکاگو به زمین نشسته و شما می خواهید در همان محدوده، شخصی را پیدا کنید که بتواند اطلاعاتی در مورد dUTPase به شما بدهد . برای این کار به ترتیب زیر عمل کنید:

۱. وارد سایت PubMed شوید.
۲. در پنجره for عبارت [AD] Chicago [TIAB] dUTPase را تایپ کنید سپس روی کلمه Go کلیک کنید.

بدین ترتیب dUTPase هم در عنوان و هم در خلاصه مقالات جستجومی شود. علاوه براین آدرس آزمایشگاه اصلی مرتبط با این مقاله هم به شیکاگو محدود می شود . با انجام جستجو چند مقاله با این خصوصیات ظاهر می شوند.

۳. روی لیست مؤلفین (که آبی رنگ است و زیر آن خط کشیده شده است) کلیک کنید.

در این حالت خلاصه مقاله و همچنین آدرس آزمایشگاه اصلی مرتبط با آن مقاله را خواهید داشت . برای این که به خلاصه همه مقالات یک جا دسترسی پیدا کنید مراحل ۵ و۴ را دنبال کنید.

۴. با کلیک روی دکمه Go Back جستجوگر خود به انتهای مرحله ۲ برگردید.
۵. در منوی مربوط به پنجره Display ، حالت نمایش را از Summary به Abstract تغییر دهید.

در این حالت لیستی از خلاصه مقالات ظاهر می شوند. در بالای هر خلاصه مقاله می توانید نام آزمایشگاه اصلی مربوط به آن مقاله را که در شیکاگو واقع است، ببینید. با استفاده از اطلاعاتی که در هر مقاله وجود دارد می توانید نام ها، آدرس خیابان و حتی آدرس پست الکترونیکی مربوط به آن مقاله را بیابید . در صورت نیاز برای تکمیل کردن اطلاعات مربوط به مقاله و پیدا کردن راهی برای تماس با افرادی که روی این مقاله کار کرده اند می توانید از یک دفتر تلفن یا یک موتور جستجو استفاده کنید. فقط از طرف ما با شماره تلفنی که پیدا کرده اید تماس نگیرید!).

اعمال محدودیت در جستجو در PubMed

PubMed راه های دیگری هم برای انجام جستجوی دقیقتر دارد. شما در یک زمینه خاص می توانید یک دامنه محدود برای جستجوی خود تعریف کنید. در ابتدای امر این کار بسیار پیچیده به نظر می رسد اما ما به شما یاد می دهیم که چگونه این کار را انجام دهید.

PMID - 9847382
OWN - NLM
STAT - MEDLINE
DA - 19990128
DCOM - 19990128
LR - 20041117
PUBM - Print
IS - 0022-538X (Print)
VI - 73
IP - 1
DP - 1999 Jan
TI - "Hidden" dUTPase sequence in human immunodeficiency virus type 1 gp120.
PG - 751-3
AB - A coding region homologous to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions of several viral lineages. ----------
-----, our results strongly suggest that an ancestral dUTPase gene has evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.
AD - Laboratory of Structural and Genetic Information, CNRS EP-91, Marseille F-12302, France. chantal@igs.cnrs-mrs.fr
FAU - Abergel, C
AU - Abergel C
FAU - Robertson, D L
AU - Robertson DL
FAU - Claverie, J M
AU - Claverie JM
LA - eng
PT - Journal Article
PL - UNITED STATES
TA - J Virol
JT - Journal of virology.
JID - 0113724
RN - 0 (HIV Envelope Protein gp120)
RN - EC 3.6.1. - (Pyrophosphatases)
RN - EC 3.6.1.23 (dUTP pyrophosphatase)
SB - IM
SB - X
MH - Amino Acid Sequence
MH - HIV Envelope Protein gp120/*chemistry
MH - HIV-1/* chemistry
MH - Humans
MH - Molecular Sequence Data
MH - Pyrophosphatases/*chemistry/genetics
MH - Research Support, Non-U.S. Gov't
EDAT - 1998/12/16
MHDA - 1998/12/16 00:01
PST - ppublish
SO - J Virol. 1999 Jan;73(1):751-3.

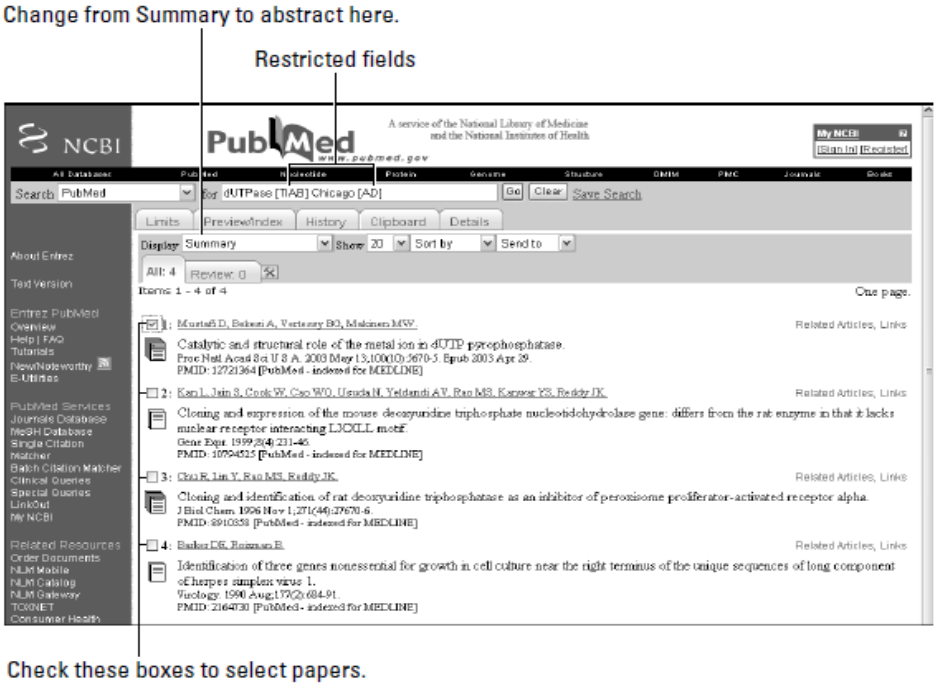
شکل ۷-۲ : ساختار داخلی یک مقاله استخراج شده از مدلاین

وقتی می خواهید یک دید کلی در مورد موضوعی که هیچ چیزی در مورد آن نمی دانید پیدا کنید در ابتدای کار به سراغ مقالات بسیار تخصصی نروید. بهتر است ابتدا به مقالات مروری[1] مراجعه کنید. در این مقالات فردی که در یک زمینه خاص کار کرده است، آخرین پیشرفت ها در آن زمینه را برای شما خلاصه کرده است. همچنین نیازی نیست که شما وقایع ۲۰ سال پیش(که در زیست شناسی مولکولی به اندازه ۳ قرن پیش است) را هم در مورد موضوع خود بدانید، مگر این که تاریخ دان باشید. بنابراین لازم است که جستجوی خود را به مقالات مروری سال های اخیر محدود کنیم. بدین منظور باید به ترتیب زیر عمل کنید:

۱. وارد سایت PubMed شوید.

[1] review articles

۲. در پنجره for عبارت dUTPase را جستجو کنید. اگر در همین حالت روی دکمه Go کلیک کنید نتیجه جستجو بیش تر از ۳۲۰ مقاله خواهد بود که برای یک فرد غیر متخصص در این زمینه خیلی زیاد است. محدود کردن جستجو به مقالات مروری سال های اخیر که به زبان انگلیسی هستند می تواند مفید باشد.



شکل ۸-۲: جستجوی متخصصین در زمینه dUTP در شیکاگو

۳. روی گزینه Limit که درست زیر پنجره Search قرار گرفته کلیک کنید. صفحه محدودیت ها ظاهر می شود (شکل ۹-۲). در زیر عبارت Limited To چندین ویژگی و زمینه برای اعمال محدودیت وجود دارد. شما می توانید بعداً به این صفحه برگردید و گزینه های مختلف را امتحان کنید.

۴. در قسمت Language گزینه English را انتخاب کنید.

۵. در قسمت Type of Article گزینه Review را انتخاب کنید.

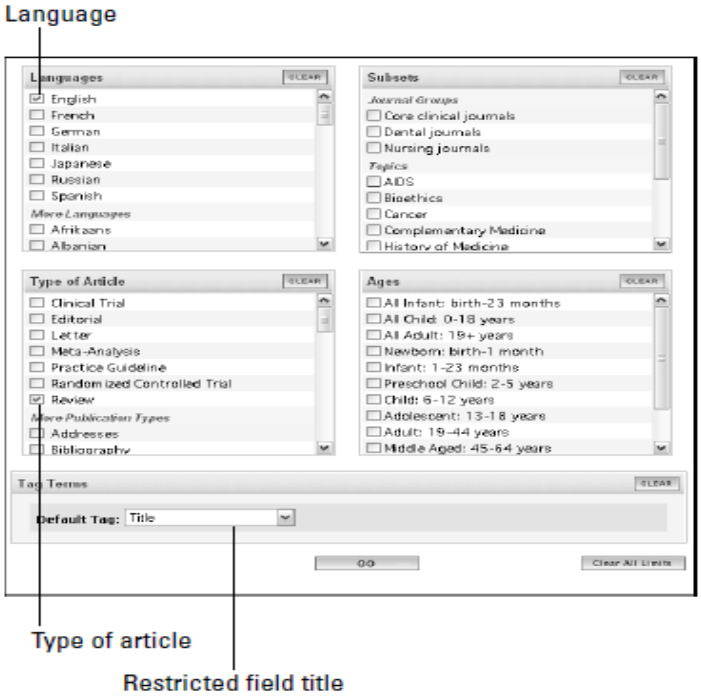
۶. در قسمت Default Tag گزینه Title را انتخاب کنید.

۷. اعمال با این محدودیت فقط مقالاتی که dUTPase در عنوان آنها هست جستجو می شوند.

۸. در نهایت روی دکمه Go کلیک کنید.

نتایج جستجو این بار با تعداد کمتری ظاهر می شوند. با اعمال این محدودیت ۵ و۴ مقاله به دست آمد که از ۲۰۰ مقاله ای که بدون اعمال محدودیت هستند که به احتمال زیاد حاوی اطلاعاتی مختصر و مفید در مورد موضوع ما هستند.

یک استفاده معمول دیگر از منوی محدودیت ها، پیدا کردن مقالاتی است که در یک بازه زمانی خاص چاپ شده اند (مثلاً جدیدترین آنها).



شکل ۹-۲: محدود کردن جستجوی کلید واژه dUTP به عنوان، مقاله مروری و زبان انگلیسی

چند نکته دیگر در مورد PubMed

برای یادگرفتن نحوه کار با PubMed به طور کامل، باید وقت زیادی صرف کرد چون PubMed موارد استفاده زیادی دارد. باید این جرأت را به خود بدهید که تک تک گزینه هایی که در این سایت وجود دارد را امتحان کنید (به خصوص آنهایی که در این کتاب در موردشان صحبت نکردیم).

نکات پایانی در مورد Pub Med:

- ✓ **چگونه تا حد ممکن مقالات نامربوط را از جستجوی خود حذف کنیم؟**
  - عباراتی را که با علامت نقل قول نوشته می شوند (مثلاً "down syndrome") مثل یک کلمه منفرد عمل می کنند و کمک می کنند تا نتیجه جستجو هرچه بیشتر با موضوع مورد نظر شما مرتبط باشد.
  - از حروف ربط AND ، OR و NOT در جای مناسب استفاده استفاده کنید.

به عنوان مثال :

DUTPase[TI] OR Pyrophosphatase[TI] NOT Smith[AU]

  - افزودن حرف اول نام مؤلف مثلاً "Abergel C " می تواند به میزان زیادی تعداد نتایج را کم کند.

- شناسه PubMed مقاله ای را که در یک زمینه خاص پیدا کرده اید (شماره ای که در قسمت PMID نوشته شده است) را یادداشت کنید. این شماره در جستجوی بعدی شما در مورد موضوعات مرتبط با موضوع قبلی (مثلاً ژن ها یا توالی های پروتئینی مرتبط) می تواند بسیار مفید باشد.
- فراموش نکنید زمانی که می خواهید یک جستجوی جدید انجام دهید، محدودیت های مربوط به جستجوی قبلی را بردارید.
- به جستجویی که هیچ نتیجه ای نداشته است فوراً اعتماد نکنید. غلط های املایی ، اعمال محدودیت های نادرست و انتخاب زمینه نامناسب می تواند یک جستجو را کاملاً خراب کند.
- وقتی به عنوان یک پژوهشگر مبتدی در مورد یک موضوع تازه جستجو می کنید، ابتدا چند خلاصه مقاله را بخوانید تا با لغات تخصصی آن موضوع آشنا شوید و سعی کنید لغات مترادف آنها را پیدا کنید. مثلاً اگر ندانید که بعضی مقالات به جای dUTPase از عبارت "dUTP Pyrophophatase " استفاده می کنند ممکن است بعضی از مقالات مرتبط را از دست بدهید.
- زمانی که نتایج جستجو کافی نیستند روی عبارت Related Articles که در سمت راست صفحه نتایج PubMed قرار گرفته کلیک کنید تا دامنه جستجوی خود را وسیع تر کنید.

✓ **مواردی که در PubMed یافت نمی شوند:**

- نام مؤلف های دهم به بعد در مقالات قدیمی تر(قبل از ۱۹۹۵)

این مسئله یک مشکل بزرگ در مورد مقالات قدیمی در زمینه ژنومیک است.

- مقالاتی که در سال های قبل از ۱۹۶۵ ثبت شده اند .

PubMed هیچ مقاله ای قبل از سال ۱۹۶۵ ندارد. اگر می خواهید یک مقاله تاریخی در رشته خودتان بنویسید، PubMed یک منبع قابل اعتماد برای شما نیست.

- خلاصه اکثر مقالاتی که در سال های قبل ۱۹۷۶ ثبت شده اند. اگر دنبال چنین مقالاتی می گردید از PubMed انتظار زیادی نداشته باشید.

بازیابی توالی های پروتئینی

جستجو های موضوعی مهم ترین استفاده ای است که زیست شناسان از PubMed در زمینه بیوانفورماتیک می کنند و بازیابی توالی پروتئین ها در درجه دوم اهمیت قرار دارد. با داشتن توالی پروتئین ها می توان اطلاعات بیشتری از موضوع مورد نظر در سطح مولکولی به دست آورد. اگر چه PubMed رابط بین شما و دنیای پروتئین ها است، ولی گاهی استفاده از آن در این زمینه مشکل است. بنابراین ما سایت دیگری را به شما معرفی می کنیم که یافتن توالی پروتئین ها در آن واقعاً ساده است.

ExPASy: یک سایت اینترنتی مناسب برای بدست آوردن اطلاعات در مورد پروتئین

سایت ExPASy که توسط یکی از پیشگامان بیوانفورماتیک پرتئین ها یعنی پروفسور Amos Bairoch طراحی شده و مدیریت می شود، بهترین منبع برای به دست آوردن اطلاعات در مورد پروتئین ها است. علاوه بر بانک اطلاعاتی Swiss–Prot که حاوی اطلاعاتی در مورد توالی پروتئین ها است، سایت ExPASy حاوی تعداد زیادی از ابزارهای آنالیزی است که ما در این کتاب از آنها استفاده می کنیم. این سایت امکان ارتباط با تعداد زیادی از سرویس دهنده های اینترنتی دیگر را نیز برای به دست آوردن آنالیزهای اختصاصی تر فراهم می کند.

بعد از به دست آوردن اطلاعات اولیه در مورد یک موضوع خاص (مثلاً dUTPase) با استفاده از PubMed، می توان پا را فراتر گذاشته و چند نمونه از توالی های پروتئینی را که فعالیت dUTPase دارند، پیدا کنید. چون اشرشیای کلی (E. coli) ارگانیسمی است که بیشترین مطالعات روی آن صورت گرفته است، به عنوان مثال چگونگی تعیین توالی پروتئینی که در این باکتری فعالیت dUTPase دارد را بررسی می کنیم :

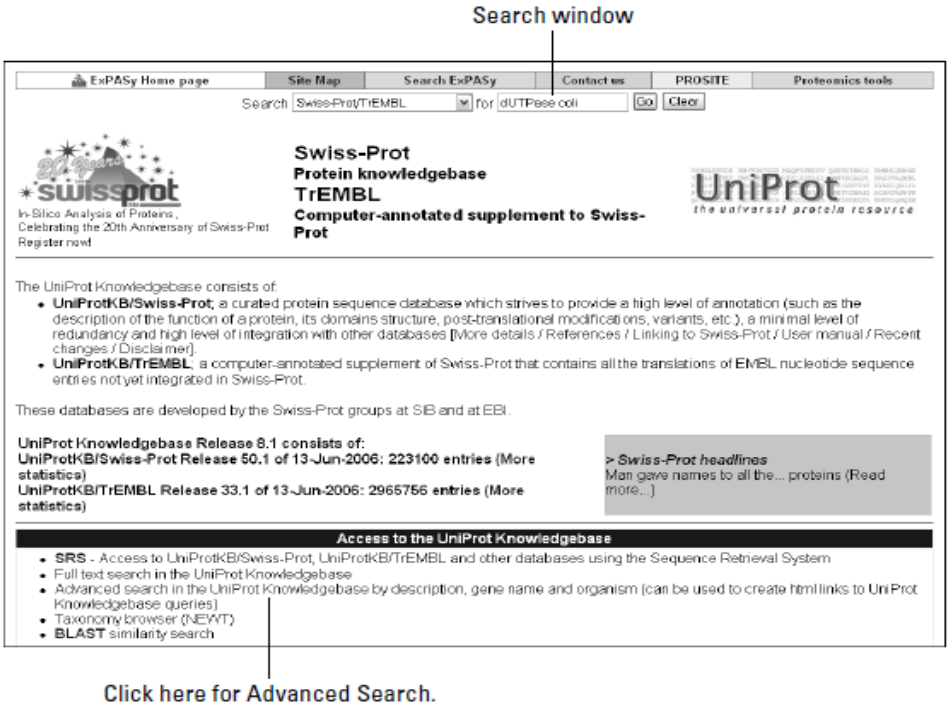
۱. وارد آدرس اینترنتی www.expasy.org/spart/ شوید. صفحه اصلی بانک اطلاعاتی Swiss – Prot باز می شود (شکل ۱۰-۲).
۲. در پنجره جستجو عبارت dUTPase Coli را تایپ کرده روی دکمه جستجو کلیک کنید.

نتیجه جستجو سه توالی پروتئینی مرتبط خواهد بود. اگر روی لینک سوم یعنی DUT ECOLI (P06968) کلیک کنید یک صفحه کامل در مورد پروتئین dUTPase اشرشیا کلی ظاهر می شود (شکل ۱۲-۲ و ۱۳-۲).

صفحه اصلی نتیجه جستجو از چهار بخش تشکیل شده است که در مورد dUTPase به ترتیب زیر است:

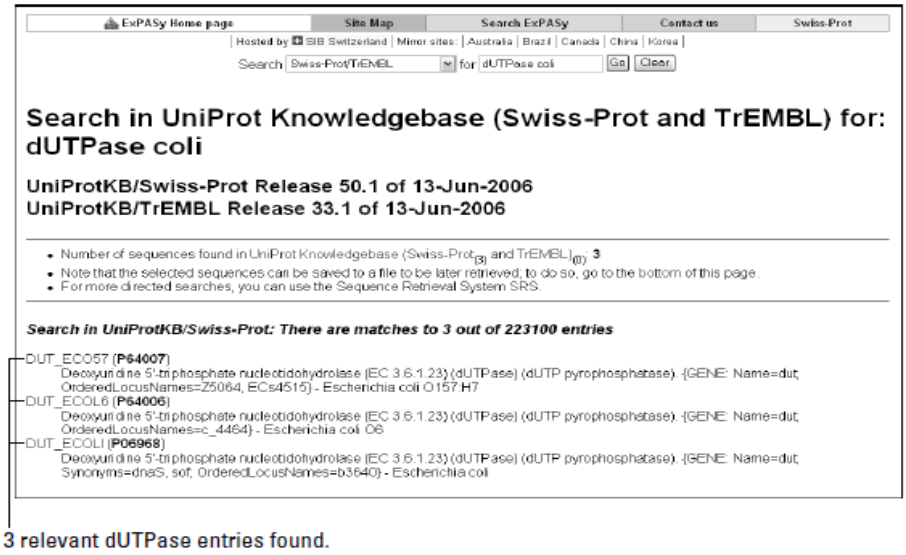
- بخش اول شامل نام این صفحه (DUT- ECOLI) در خط اول ، و یک شناسه منحصر به فرد (P06968) که" شماره مسلسل اولیه[1] " نامیده می شود، در خط دوم است. مانند شناسه ی PubMed، این شماره ها هم ارزش یادداشت برداری را دارند. این شماره ها برای مراجعه به اطلاعات مرتبط در بانک های اطلاعاتی دیگر مفیدند .
- بخش دوم شامل توضیحات بیوشیمیایی پروتئین شامل نام استاندارد، شماره کمیته بین المللی آنزیم (E.C) و همچنین تعدادی واژه معادل برای آن پروتئین است. این واژه های معادل می توانند جستجوی شما را در مدلاین وسیع تر کنند. همچنین در این بخش لیستی از منابع موضوعی مرتبط با توالی یابی این پروتئین با بعضی مطالعات کاربردی آمده است. توجه کنید که " شناسه های PubMed " همه مقالاتی که در این قسمت به آنها اشاره می شود ذکر می گردند تا مراجعه متقابل امکان پذیر باشد.

[1] primary accession number

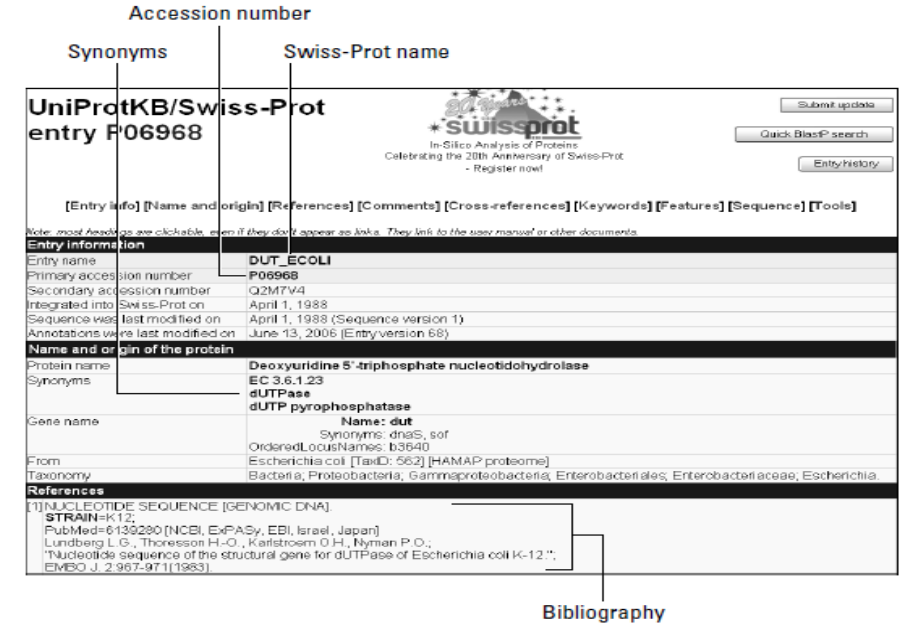


شکل ۱۰-۲: پنجره جستجو در بالای صفحه اصلی سایت Swiss-Prot

- بخش های سوم شامل سری های کاملی از پیوند ها (Links) با برنامه های مختلفی است که دسته بندی های کاربردی انجام می دهند، از جمله نواحی پروتئینی مرتبط، ساختارهای سه بعدی و گروه های عاملی (در فصل ۴ اطلاعات بیشتری در این مورد وجود دارد).
- بخش چهارم شامل توالی واقعی آمینواسیدی پروتئین است. در شکل ۱۴-۲ می توانید ببینید که dUTPase باکتری *E. coli* از ۱۵۱ اسید آمینه تشکیل شده است و وزن مولکولی آن ۱۶۱۵۵ Da می باشد. در این شکل توالی آمینواسیدی به صورت کدهای یک حرفی نشان داده شده است (برای اطلاعات بیشتر در مورد فرم های نام گذاری آمینواسیدها در فصل ۱ آمده است.). برای انجام مطالعات بیشتر لازم است که توالی های آمینواسیدی به فرم FASTA باشند. این فرم نامگذاری آمینواسیدها بسیار ساده تر است و با علائم نوشتاری هم تداخلی ندارد. برای اطلاعات بیشتر در این مورد می توانید به بخش " The FASTA Land RAWL format " مراجعه کنید.



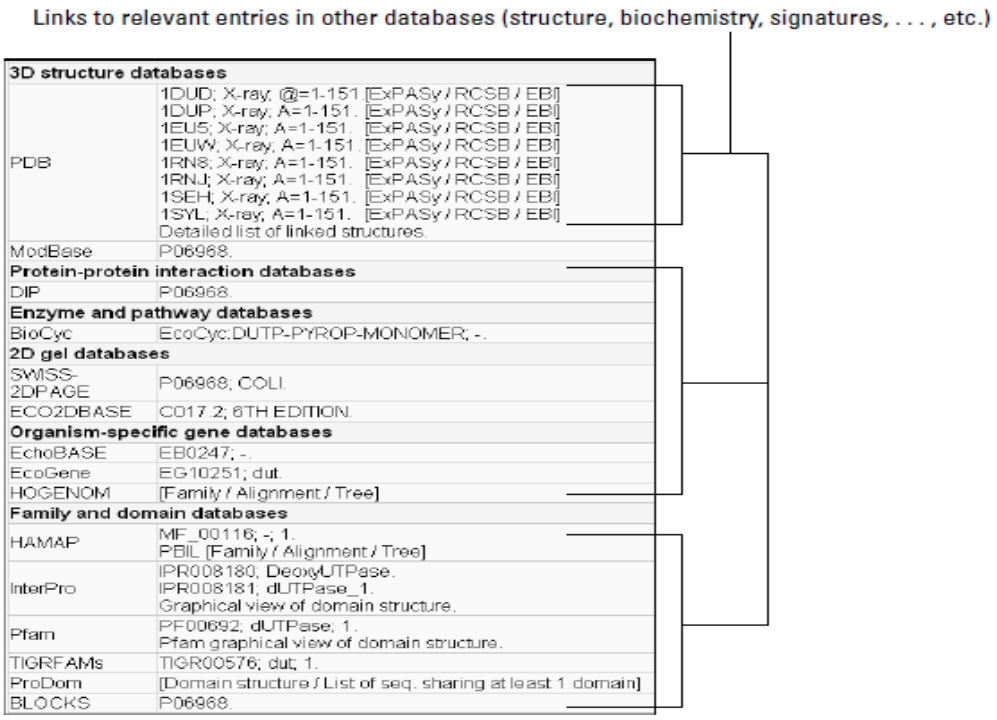
شکل ۱۱-۲: سه توالی dUTPase از باکتری *E. coli*



شکل ۱۲-۲: نتیجه جستجوی پروتئین dUTPase باکتری *E. coli* در بانک اطلاعاتی Swiss-Prot (بخش بالایی)

۳. برای تبدیل توالی آمینواسیدی به فرم FASTA روی گزینه FASTA format کلیک کنید. فرم FASTA توالی شما در شکل ۱۵-۲ نشان داده شده است.

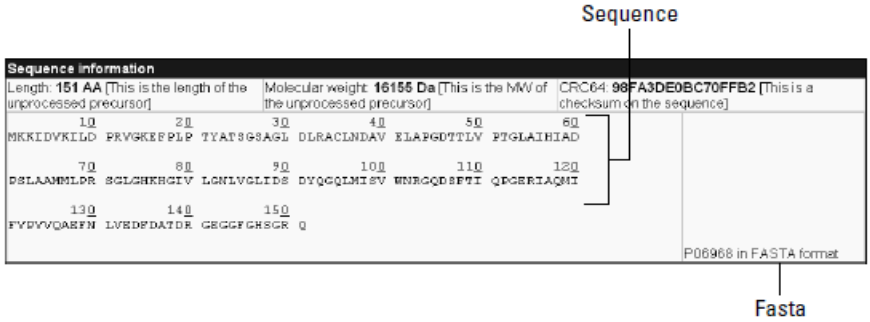
با انجام این سه مرحله شما به آسانی توانستید از یک واژه مثل dUTPase به یک تعریف کامل از پروتئینی که این عمل را (dUTPase) در همه انواع باکتری ها انجام می دهد برسید (اکنون شاید تصور کنید که به آخر بیوانفورماتیک رسیده اید!!).



شکل ۱۳-۲: نتیجه جستجوی پروتئین dUTPase باکتری *E. coli* در بانک اطلاعاتی Swiss-Prot (بخش میانی)

راه های پیشرفته تر برای بازیابی توالی پروتئین ها

تشخیص صحیح پروتئین در Swiss – Prot همیشه این قدر ساده نیست چون ممکن است اطلاعات اولیه ای که با آن جستجو را شروع می کنید، به اندازه کافی اختصاصی نباشد. مثلاً اگر کلمه ای که با آن جستجو می کنیم در چند بخش از Swiss – Prot یافت شود ممکن است نتیجه جستجوی ما زیاد مطلوب نباشد . برای بدست آوردن نتایج بهتر شما باید زمینه جستجوی فرد را محدود کنید (در مورد چگونگی انجام این کار در PubMed بحث شد). این نوع جستجو محدود شده، به تشکیلات هر بانک اطلاعاتی و همچنین ویژگی های نرم افزارهای جستجو بستگی دارد.



شکل ۱۴-۲: توالی آمینو اسیدی پروتئین dUDPase باکتری E. coli (بخش پایینی)

```
>sp | P06968 | DUT_ECOLI Deoxyuridine 5' -triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAPGDTTLVPTGLAIHIAD
PSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWNRGQDSFTIQPGERIAQMI
FVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ
```

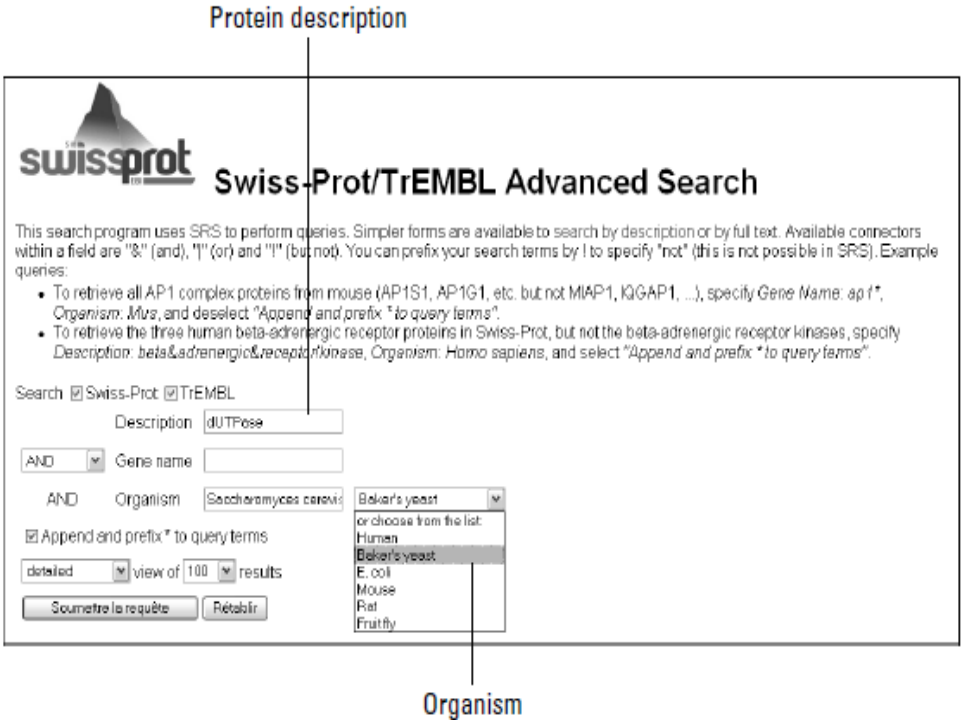
شکل ۱۵-۲: توالی پروتئین dUTPase باکتری E. coli با فرمت FASTA

در Swiss – Prot برای محدود کردن جستجو در یک زمینه خاص به ترتیب زیر عمل کنید:

۱- وارد آدرس اینترنتی www.expasy.org/spart/ شوید. صفحه اصلی بانک اطلاعاتی Swiss – Prot ظاهر می شود. (شکل ۱۰-۲)

۲- روی عبارت Advanced Search in the UniProt Knowledgebase (که زیر نوار Access to the UniProt Knowledgebase قرار گرفته است) کلیک کنید.

۳- صفحه جستجوی پیشرفته ظاهر می شود (شکل ۱۶-۲). شما در این صفحه الکترونیک می توانید با وارد کردن لغات کلیدی مرتبط با سه زمینه "Description"، " Gene name" و " Organism" در بخش مربوطه، توالی پروتئین مورد نظر خود را بیابید.



شکل ۱۶-۲: استفاده از جستجوی پیشرفته در بانک اطلاعاتی Swiss-Prot

در مورد مثالی که تاکنون دنبال کردیم شما باید عبارت dUTPase را در پنجره Description تایپ کنید، از منوی مرتبط با پنجره Organism گزینه baker's yeast را انتخاب کنید و سپس روی کلمه Submit Query کلیک کنید (شکل ۱۶-۲). نتایج جستجوی شما در صفحه ای جداگانه ظاهر می شوند و این نتایج دقیقاً با پروتئین dUTPase مخمر مورد نظر شما مرتبط هستند. نام این ورودی DUT-YEAST وشماره مسلسل آن P33317 است.

فرمت FASTA و RAW

FASTA نام یک برنامه پرطرفدار تطبیق توالی و کاوشگر بانک های اطلاعاتی است که در سال ۱۹۸۸ توسط W.R.Pearson, D.J.Lipman ساخته شده است. (با روش های جستجویی که در PubMed یادگرفته اید الان می توانید با استفاده از این دو نام ، مقاله اصلی را پیدا کنید.)

FASTA توالی ها را به فرم زیر نمایش می دهد:

> My- Sequence-Name
> ARCGTCRGCKINTANDRGCKINTAND
> CKINTANDARCGTCRGCKINTANDRG
> CKINTAND…

خط اول که با علامت " > " شروع می شود حاوی شناسه توالی و یک تعریف کوتاه اختیاری است (خط تعریف)[1] . خطوط بعدی توالی DNA یا پروتئین را به صورت کدهای یک حرفی نشان می دهند. این توالی تا علامت ">" بعدی ادامه دارد. فرم FASTA، به دلیل سهولت تحلیل آن، بسیار پرطرفدار و فرم قراردادی نمایش توالی ها در بسیاری از نرم افزارهای آنالیزگر توالی ها از جمله BLAST و CLUSTALW است. اما توجه کنید که برنامه هایی که از فرمت FASTA استفاده می کنند، گاهی به کوچک یا بزرگ بودن حروف انگلیسی حساس (Case-sensitive) هستند. نکات زیر باید رعایت رعایت شود:

- ✓ همواره برای نمایش توالی به صورت کدهای یک حرفی، از حروف بزرگ الفبا استفاده کنید.
- ✓ هنگامی که در کامپیوتر شخصی خود از توالی های به فرم FASTA استفاده می کنید نرم افزار پردازش کلمه خود را درحالت TEXT قرار دهید . با این کار از فرمت رهایی می یابید و تنها از کاراکترهای ASCII استفاده می کنید.
- ✓ برای نمایش این توالی ها در یک برنامه پردازش کلمه برای راحت تر شدن عمل تطبیق از قلم Courier استفاده کنید.

در بعضی برنامه های آنالیز توالی از فرم RAW استفاده می شود. تنها تفاوت این فرم با فرم FASTA این است که فرم RAW خط تعریف را ندارد (اما ابزارهای آنالیز ممکن است خیلی سخت گیر باشند). زمانیکه برنامه آنالیز توالی به فرم RAW نیاز داشته باشد و شما فرم FASTA را به آن بدهید ممکن است باعث ایجاد خطا شود یا ممکن است تصادفاً تعریف یک توالی شبیه یک توالی پروتئین یا DNA باشد!

بدست آوردن توالی های پروتئینی مرتبط با هم

در بسیاری از تحقیقات در زمینه زیست شناسی مولکولی نیاز به پیدا کردن مجموعه ای از توالی های پروتئینی با عملکرد مشابه داریم. کار اصلی در این تحقیقات تشخیص قطعات عملکردی محافظت شده (قطعاتی از توالی ها که در پروتئین های مختلف مشابه به نظر می رسند و عملکرد مشابهی هم دارند)، تطبیق همزمان چندین توالی و ارزیابی تغییر پذیری آنها با مطالعات فیلوژنتیکی (ارتباط تکاملی بین توالی ها) است.

[1] - the definition line

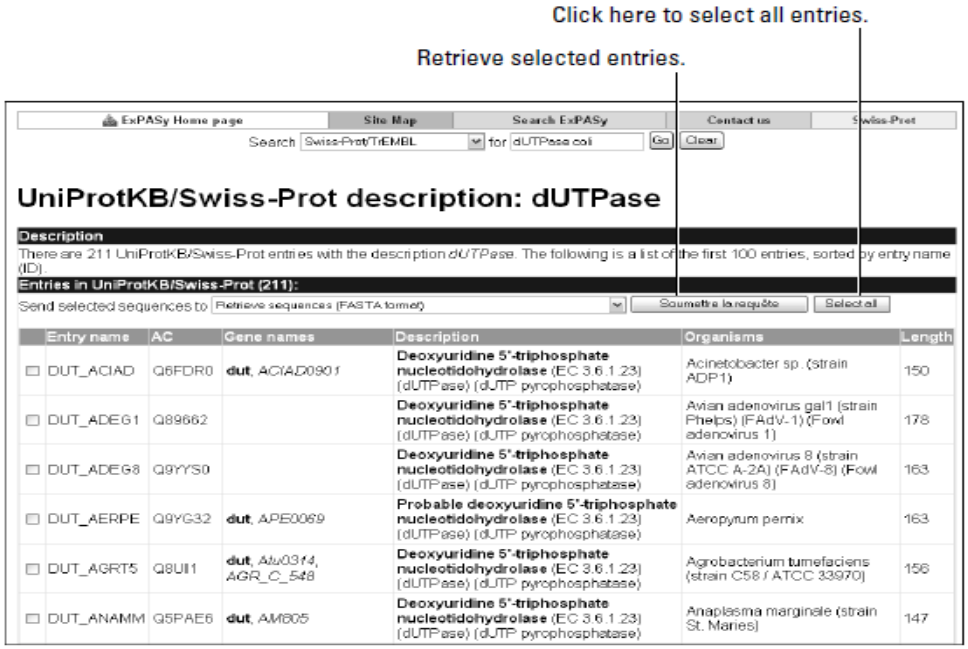
اگر یک فایل شخصی از توالی ها در کامپیوتر خود بسازید، می توانید در دیگر برنامه های آنالیز که در سایت های دیگر موجودند از این اطلاعات ذخیره شده استفاده کنید. اکنون با استفاده از مثال dUTPase به شما نشان خواهیم داد که چگونه فایلی از همه توالی های پروتئینی مرتبط با dUTPase بسازید. برای اینکه این فایل در برنامه های دیگر هم قابل استفاده باشد، آن را به فرم FASTA درست می کنیم.

مراحل انجام این کار به ترتیب زیر است:

۱. وارد آدرس اینترنتی www.expasy.org/sprat/ شوید و روی گزینه advanced search in the Uniprot knowledgebase کلیک کنید.
۲. در پنجره جستجو (درست بالای پنجره Description) ، Swiss-Prot را در حالت انتخاب باقی بگذارید، ولی TrEMBL را از حالت انتخاب خارج کنید (شکل ۱۶-۲).
TrEMBL یک بانک اطلاعاتی متشکل از ترجمه توالی های جدید DNA است که نظارتی روی آن نیست، اما Swiss-Prot فقط شامل ورودی های اطلاعات است که اعتبار آنها توسط افراد متخصص تأیید شده است. براساس آخرین اطلاعات ما با محدود کردن جستجو به Swiss-Prot شما می توانید مطمئن باشید همه پروتئین هایی که در این جستجو یافت می شوند dUTPase واقعی هستند.
۳. در پنجره Description عبارت dUTPase را تایپ کنید.
دقت کنید چیز دیگری تایپ نکنید و بخصوص مطمئن شوید که در پنجره Organism چیزی نوشته نشده باشد.
۴. روی دکمه Submit Query کلیک کنید.
صفحه نتایج ظاهر می شود (شکل ۱۷-۲). زمانیکه ما در حال نگارش این کتاب بودیم با انجام این جستجو ۲۱۱ ورودی Swiss-Prot حاصل شد. اگر شما این جستجو را انجام دهید ممکن است نتایج بیشتری حاصل شود. با مرور لیست حاصل متوجه خواهید شد که همه ورودی ها عالی به نظر می رسند و بخوبی توالی های پروتئین dUTPase را نشان می دهند. با کلیک بر روی هر کدام از نام ها (مثلاً DUT-ADEG1) می توانید به «کارت شناسایی» مربوطه دسترسی پیدا کنید. برای اینکه بدانید در این کارت های شناسایی چه اطلاعاتی یافت می شود، می توانید به شکل های ۱۲-۲، ۱۳-۲ و ۱۴-۲ مراجعه کنید.

اما برای اینکه کار جمع آوری اطلاعات در مورد مثال dUTPase را به پایان برسانید باید همه این توالی ها را در یک فایل جمع کنید. پس مراحل زیر را ادامه دهید:

۱. بعد از انجام مراحل ۱ تا ۴ که در بالا ذکر شد روی دکمه Select All واقع در بالای سمت راست صفحه نتایج، یا روی مربع کوچک کنار توالی های خاصی که مورد علاقه شماست کلیک کنید (شکل ۱۷-۲).



شکل ۱۷-۲: قسمت بالای لیست پروتئین های dUTPase یافت شده در بانک اطلاعاتی Swiss-Prot

۲. روی کلمه Send Query کلیک کنید.

صفحه دانلود ظاهر می شود (شکل ۱۸-۲). توالی های شما اکنون برای ذخیره شدن آماده هستند.

محتویات این صفحه را بسته به اینکه از چه نوع مرورگر اینترنت استفاده می کنید، می توانید به روش های مختلفی ذخیره کنید. ما انجام روند زیر را به شما پیشنهاد می کنیم، چون همیشه عملی است و تغییری در فایل های حاصل ایجاد نمی کند.

۱. در منوی Edit ابتدا گزینه Select All و سپس در همین منو گزینه Copy را انتخاب کنید.
۲. یک محیط جدید برنامه پردازش گر کلمات باز کنید (مثلاً محیط Microsoft Word).
۳. در این محیط جدید از منوی Edit گزینه Paste را انتخاب کنید.
۴. همه اطلاعات منتقل شده را دوباره روی قلم courier و اندازه ۱۰ یا ۸ تنظیم کنید.
۵. در نهایت اطلاعات خود را با نام dUTPaseDB.txt و با انتخاب(*.txt) در قسمت Save as type ذخیره کنید.

اکنون شما همه توالی های شناخته شده پروتئین dUTPase را در کامپیوتر شخصی خود دارید. این اطلاعات را نگه دارید چون در بخش های بعدی این فصل به آنها نیاز خواهید داشت.

شکل ۱۸-۲: توالی های FASTA که آماده برای ذخیره شدن هستند

بازیابی توالی های DNA

توالی های پروتئینی ساده هستند، اندازه آنها در دامنه باریکی قرار می گیرد (به استثنای تعداد کمی پروتئین غول پیکر، اندازه آنها در محدوده ۲۰۰ ± ۳۰۰ آمینواسید قرار می گیرد)، حد و مرز آنها کاملاً مشخص است و ویژگی های کاربردی اختصاصی دارند. بعلاوه، پروتئین های میکروب ها و یوکاریوتهای متکامل تر (حیوانات و گیاهان) تقریباً ویژگی های مشابهی دارند.

اما همانطور که انتظار می رود توالی های DNA مربوط به پروتئین ها در حیوانات تکامل یافته، متنوع تر و پیچیده تر می شوند. اندازه ژن ها در انسان ممکن است از چند هزار جفت باز (هم اندازه با ژن های میکروب ها) تا چند صد هزار جفت باز متغیر باشد.

همه بخش های مولکول DNA به پروتئین کد نمی شود

انواع مختلفی از توالی های DNA یک ژن را تشکیل می دهند:

- ✓ نواحی تنظیمی (معمولاً قبل از ناحیه کد کننده)
- ✓ نواحی که ترجمه نمی شوند. این نواحی قبل و بعد از نواحی کد کننده واقع می شوند.
- ✓ نواحی کد کننده پروتئین

در یوکاریوت ها (مخمر ها، گیاهان و حیوانات)، ناحیه کد کننده پروتئین به تعداد متغیری «اگزون[1]» تقسیم می شود که توسط قطعات «اینترون» از هم جدا می شوند. اگزون ها قطعاتی هستند که در تولید پروتئین نهایی شرکت می کنند ولی اینترونها در

[1] exon

این کار شرکت نمی کنند.در نتیجه کار کردن با توالی های DNA همواره سخت تر از کار کردن با توالی های پروتئین است. در فصل های ۳ و ۵ توضیحات بیشتری در مورد ساختار ژن ها خواهید دید.

چگونه هنگام دانلود فایل های توالی ها، از کاراکترهای مزاحم در امان باشیم؟

در مثال dUTPase ما برای ذخیره کردن اطلاعات بدست آمده از مسیر File → Save As استفاده نکردیم، چون فایلی که با این روش ذخیره می شود مشکلاتی دارد:

- ✓ این فایل ها ممکن است حاوی کاراکترهای مزاحم مخفی باشند (control/Alt/shift+sth) که معمولاً بصورت <PRE> در ابتدای فایل ظاهر می شوند و می توانند فرم FASTA را به هم بریزند.
- ✓ این فایل ها نمی توانند توسط برنامه پردازش کلمه کامپیوتر شما دوباره باز شوند چون پسوند مناسبی ندارند (مثلاً doc.) یا ممکن است از شما درخواست کند که یک برنامه رمزگذاری نامفهوم را برای بازکردن فایل انتخاب کنید که ما در مورد آن چیزی نمی دانیم و استفاده از آن را به شما هم توصیه نمی کنیم.

تجربه ما می گوید استفاده از مسیر File → Save As نتایج غیر قابل پیش بینی ایجاد می کند که به نوع، نسخه و نحوه کار مرورگر اینترنت شما بستگی دارد. بنابراین بعنوان یک تمرین می توانید بررسی کنید که آیا در ابتدا یا انتهای توالی های بدست آمده از اینترنت علائم غیر منتظره ای وجود دارد یا نه ؟! در اکثر برنامه های آنالیز توالی، توالی های به فرم FASTA باید با «خط تعریف» شروع شوند و نه چیز دیگری، مثلاً :

 P0343456
My-Sequence-definition

هر کاراکتر ابتدایی که بجای < بنشیند (حتی اگر هیچ علامتی جای آن ننشیند)، می تواند موجب ایجاد خطا شود (مثلاً ممکن است خط تعریف، بعنوان بخشی از توالی پروتئین شناخته شود). همچنین هر توالی باید با کاراکتر <New Line> به پایان برسد (این کاراکتر انتهایی نقش یک خط سفید را بازی می کند).

هرچند معمولاً هیچ محدودیتی برای طول خط تعریف یا قسمت توالی وجود ندارد، اما برای ایمنی بیشتر اطلاعات خود بهتر است از تعداد معقولی کاراکتر (کمتر از ۱۰۰-۶۰) استفاده کنید. باید توجه شود که اولین کاراکتر خط تعریف باید < باشد ولی بعد از این کاراکتر می توان حتی از کاراکترهای تهی مثل <spase> یا <tab> استفاده کنید، بدون اینکه در تجزیه توالی ها اختلالی ایجاد شود.

و نکته آخر اینکه از کاراکترهایی که از جمله کدهای استاندارد آمینو اسید ها نیستند مثل <-> یا <*> استفاده نکنید چون برنامه های آنالیز مختلف، رفتار مشابهی با این کاراکترها ندارند. این کاراکترها یا از قلم می افتند (یک محل آمینو اسیدی حذف می شود)، یا با X جایگزین می شوند یا به سادگی ممکن است ایجاد خطا کنند (برای اطلاعات بیشتر در مورد کدهای استاندارد آمینو اسیدها به فصل ۱ مراجعه کنید).

دستیابی به توالی DNA از روی توالی پروتئین

در بانک های اطلاعاتی ارتباط بین توالی های پروتئین و DNA یک ارتباط یک به یک نیست. بسیاری از توالی های DNA کاملاً متفاوت که گاهی حتی هیچ همپوشانی با هم ندارند، ممکن است مربوط به یک پروتئین یا یک ژن خاص باشند. مثلاً همه موارد زیر می توانند مربوط به یک ژن یا پروتئین خاص باشند (همچنین شکل های ۱-۳ و ۲-۳ را ببینید):

- ✓ رونوشت اولیه[1] که کپی توالی DNA یا ژن، از ابتدا تا انتهاست و هم اگزون دارد و هم اینترون.
- ✓ رونوشت بالغ[2] که مولکول mRNA ای است که از حذف اینترون ها از رونوشت اولیه حاصل می شود.
- ✓ منطقه کد کننده پروتئین که «قالب خواندن باز» یا ORF هم نامیده می شود.
- ✓ انواع زیادی توالی های ناکامل

اگر صرفاً کارکردن با توالی های پروتئین ها کافی بود زندگی بسیار ساده می شد، اما گاهی شرایطی پیش می آید که مجبور می شوید به توالی DNA برگردید. این شرایط بیشتر زمانی پیش می آید که می خواهید یک ژن را (با استفاده از تکنیک PCR[3]) تکثیر کرده، آن را به یک ارگانیسم دیگر منتقل کنید و سپس پروتئین مربوطه را در آن ارگانیسم تولید کنید یا یک جهش خاص را در آن ژن القا کنید. مشکلی که وجود دارد اینست که چگونه با داشن یک توالی پروتئینی می توان توالی ناحیه کد کننده این پروتئین را روی DNA پیدا کرد؟

یافتن توالی DNA مربوط به یک پروتئین خاص

فرض کنید می خواهید توالی DNA لازم برای کلون کردن ژن dUTPase باکتری *E. coli* را پیدا کنید. به نظر مشکل می رسد، اما با دنبال کردن مراحل زیر می توانید این کار را انجام دهید:

۱. وارد آدرس اینترنتی www.expasy.org/sprot/ شوید.

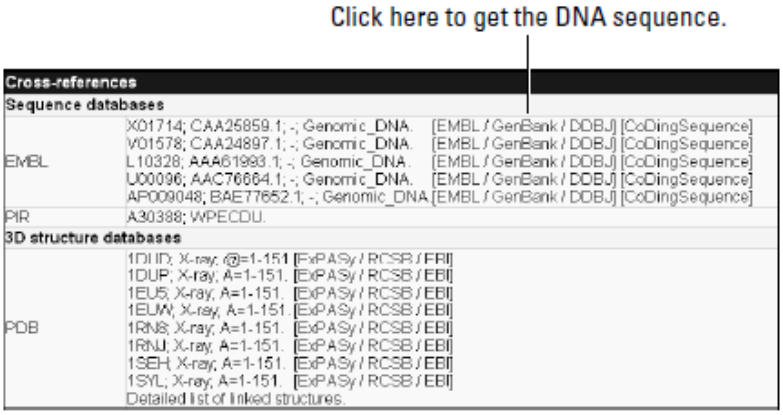
۲. برای دستیابی سریع به ورودی E.coli dUTPase کافی است شماره مسلسل مربوطه (P06968) را در پنجره جستجوی واقعی در بالای صفحه نوشته و روی گزینه search کلیک کنید. بدین ترتیب براحتی به صفحه اطلاعات مربوط به پروتئین P06968 دسترسی پیدا می کنید (شکل ۱۲-۲).

۳. روی لینک Cross-References که تقریباً در بالای صفحه قرار دارد کلیک کنید.

مرورگر شما به قسمت مربوط وارد می شود (شکل ۱۹-۲). بخش Cross-References حاوی چند قسمت است که هر کدام با یک کلمه کلیدی مثل EMBL , PIR , PDB و ... مشخص شده است. این لغات کلیدی مربوط به انواع مختلفی بانک های اطلاعاتی هستند که اطلاعات بیشتری در مورد پروتئین شما در اختیارتان قرار می دهند. با کلیک بر روی هر کدام از لغاتی که زیر آنها خط کشیده شده است به ورودی مربوطه در بانک اطلاعاتی مربوطه خواهید رفت. در

[1] primary transcript
[2] mature transcript
[3] polymerase chain reaction

واقع این بخش مجموعه ای از راه های ارتباطی بین سایت های مختلف است، که همه آنها با یک موضوع مشخص مرتبط هستند. بعنوان مثال قسمت EMBL را در نظر می گیریم. در این قسمت هر خط با یک شماره نامفهوم شروع می شود.
بعد از این شماره ها راه های ارتباطی با EMBL , Gen Bank , DDBJ (که بانک های اطلاعاتی هستند) و Coding sequence (که از طریق آن می توان دقیقاً به قطعه ای از DNA که کدکننده پروتئین است دسترسی پیدا کرد) قرار گرفته اند. پس در این بخش ۴ راه ارتباطی وجود دارد. اکنون اجازه بدهید ببینیم وقتی روی آنها کلیک کنیم چه اتفاقی می افتد.



شکل ۱۹-۲: بخش cross-references از صفحه اطلاعاتی P06968

۴. روی لینک GenBank در اولین خط ردیف EMBL کلیک کنید.

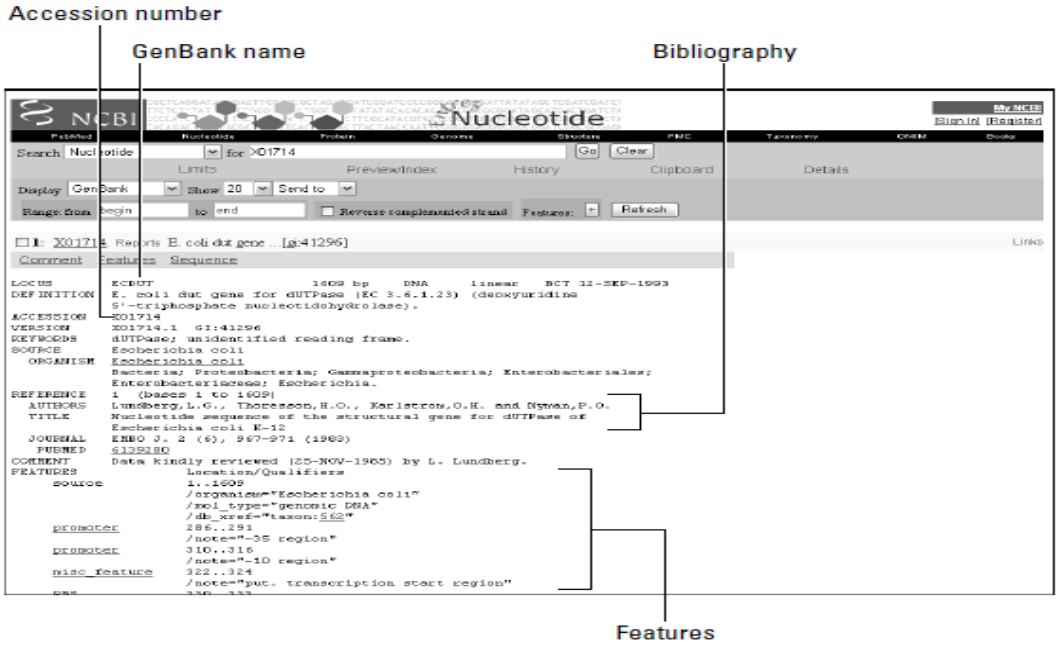
اطلاعات مرتبط با پروتئین P06968 در بانک اطلاعاتی Gen Bank (با شماره مسلسل X01714) در چندین صفحه ظاهر می شود (شکل ۲۰-۲).

بطور کلی یک ورودی GenBank چهار بخش دارد و این ۴ بخش برای مثال dUTPase به صورت زیر است:

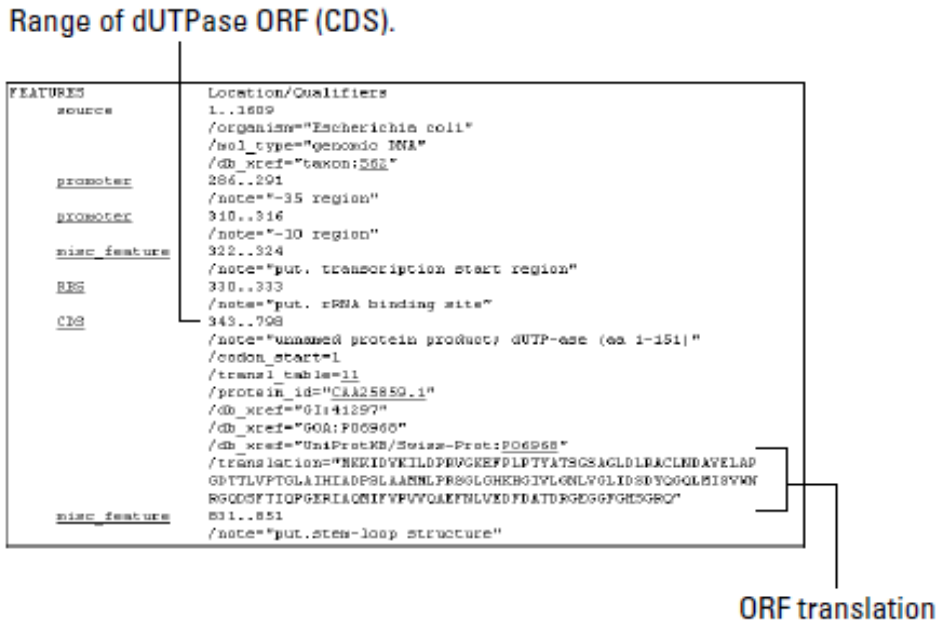
- Locus name (ECDUT) که یک شناسه قراردادی است و پس از آن یک تعریف کوتاه و یک شماره مسلسل منحصر به فرد (X01714) قرار گرفته اند. این شماره بهترین شناسه برای این ورودی است و بهتر است آن را یادداشت کنید، چون هنگام مراجعه به ورودی های دیگر در بانک های اطلاعاتی دیگر می توانید از آن استفاده کنید (کاری که در Swiss-Prot هم انجام دادیم). چند خط پایین تر در قسمت ORGANISM , SOURCE توضیحی در مورد منشأ بیولوژیک توالی DNA مربوطه آمده است.
- در بخش Reference مقالاتی که مرتبط با تعیین این توالی هستند فهرست شده اند. این فهرست در مورد توالی های بزرگ می تواند خیلی طولانی شود.
- بخش Features حاوی تعریف و محدوده دقیق عوامل مختلفی است که در توالی DNA تشخیص داده شده اند. در مثال ما (ورودی X01714) این قسمت شامل عوامل پروموتور[1]، محل های اتصال ریبوزوم[2] (RBS) و قطعات

[1] promoter
[2] ribosome binding site

کد کننده پروتئین[1] (CDS) است. در شکل ۲۱-۲ درست در مقابل قسمت CDS : محدوده کد کننده پروتئین dUTPase بصورت ۷۹۸ ... ۳۴۳ و در زیر این محدوده، نام و ترجمه آمینو اسیدی پروتئین مربوطه نشان داده شده است.



شکل ۲۰-۲: بخش بالایی ورودی ECDUT/X01714 در بانک اطلاعاتی GeneBank

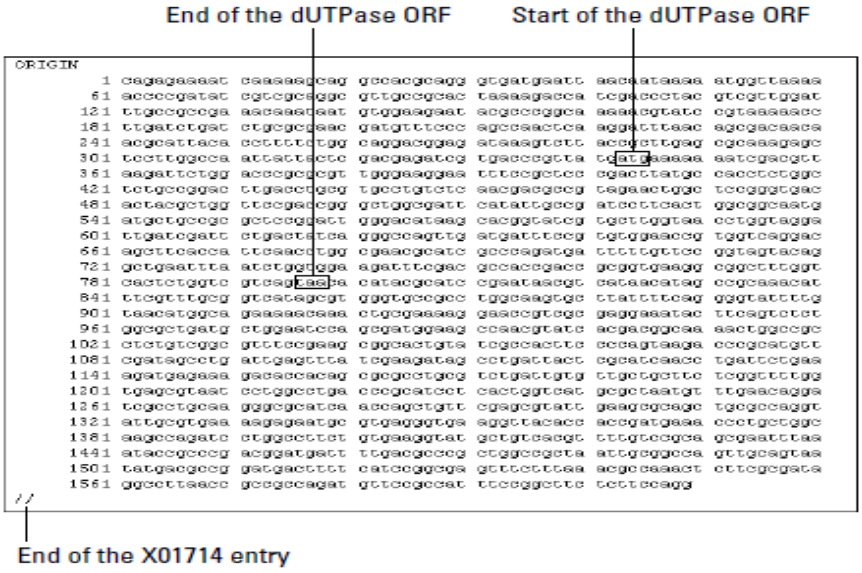


شکل ۲۱-۲: بخش Features ورودی ECDUT/X01714 بانک اطلاعاتی GenBank

[1] protein coding segments

- در بخش Sequence (بخش انتهایی ورودی GenBank)، نوکلئوتیدها بین کلید واژه ابتدایی و علامت // که نشاندهنده اتمام توالی است، نوشته شده اند. شماره هایی که در کنار توالی نوشته می شود، یافتن محل دقیق ناحیه کد کننده dUTPase را راحت تر می کند (شکل ۲۲-۲).

با استفاده از روشی که قبلاً ذکر شد می توانید همه این اطلاعات را در کامپیوتر خود ذخیره کنید.

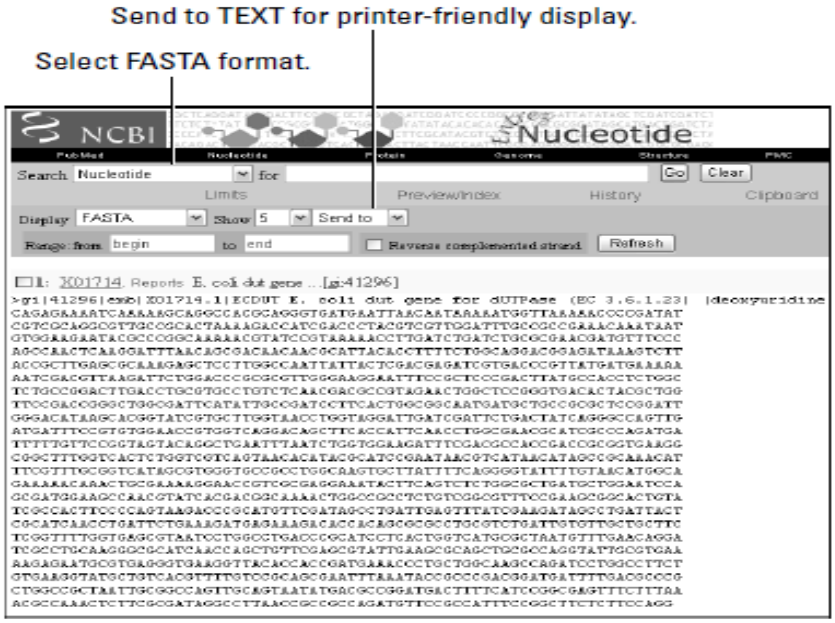


شکل ۲۲-۲ : بخش sequence ورودی ECDUT/X01714 بانک اطلاعاتی GeneBank

برای اینکه از این توالی نوکلئوتیدی بتوان بعنوان اطلاعات اولیه برای استفاده از برنامه های دیگر (مثلاً طراحی پرایمر) استفاده کرد، ممکن است لازم باشد که آن را از قسمت متن ورودی جدا و آن را به فرم FASTA تبدیل کرد. برای این کار:

۱. به بالای صفحه ورودی ECDUT/X01714 برگردید. صفحه کامپیوتر شما باید چیزی شبیه شکل ۲۰-۲ باشد.
۲. از منوی باز شونده به پایین Display فرم FASTA را انتخاب کنید (شکل ۲۳-۲).
۳. با انتخاب گزینه Text از پنجره ای که در حالت عادی درون آن عبارت sent to نوشته شده متن را به حالت قابل فهم تری تبدیل کنید.
۴. با انجام مراحل زیر توالی به فرم FASTA را ذخیره کنید:

أ- در منوی Edit مرورگر اینترنت خود ابتدا گزینه Select All و سپس گزینه copy را انتخاب کنید.
ب- یک فایل Word باز کنید و در منوی Edit آن روی گزینه Paste کلیک کنید. سپس قلم courier و اندازه قلم ۸ یا ۱۰ را انتخاب کنید.
ت- با قرار دادن save as type در حالت text only (*.txt) اطلاعات خود را با نام dUTPase DNA.txt ذخیره کنید.



شکل ۲۳-۲: توالی نوکلئوتیدی ورودی ECDUT/X01714 در GenBank به فرم FASTA

مقایسه یک توالی پروتئین با توالی های دیگر با استفاده از BLAST

بعد از اینکه مبانی بازیابی توالی های پروتئین از بانک های اطلاعاتی را یاد گرفتید، اکنون می توانید اولین آنالیز خود را روی یک توالی پروتئین انجام دهید. برای انجام این کار بهتر است یک جستجو در BLAST انجام دهیم. BLAST [1] یک ابزار مقایسه کننده توالی است که می تواند به سرعت پروتئین هایی را که توالی شبیه به پروتئین شما دارند را پیدا کند. این اطلاعات می توانند برای اهداف مختلف از جمله پیش بینی نقش عملکردی پروتئین، ساختار سه بعدی و نحوه سازماندهی نواحی مختلف پروتئین یا شناسایی پروتئین های مشابه در سایر ارگانیسم ها مورد استفاده قرار گیرند (در فصل ۷ جزئیات چگونگی استفاده از BLAST شرح داده خواهد شد. در این بخش فقط یاد می گیریم که چگونه کار کردن با BLAST را شروع کنیم.)

BLAST تحت حمایت مرکز NCBI [2] (که از قبل بعنوان مرکز حمایت کننده بانک اطلاعاتی موضوعی PubMed با آن آشنا هستید) است. بزودی متوجه می شوید که استفاده از BLAST هم مثل استفاده از PubMed ساده است. برای شروع مراحل زیر را انجام دهید:

۱. وارد آدرس اینترنتی www.ncbi.nlm.nih.gov/BLAST/ شوید.

[1] Basic local aligment search tool
[2] National center for biotechnology information

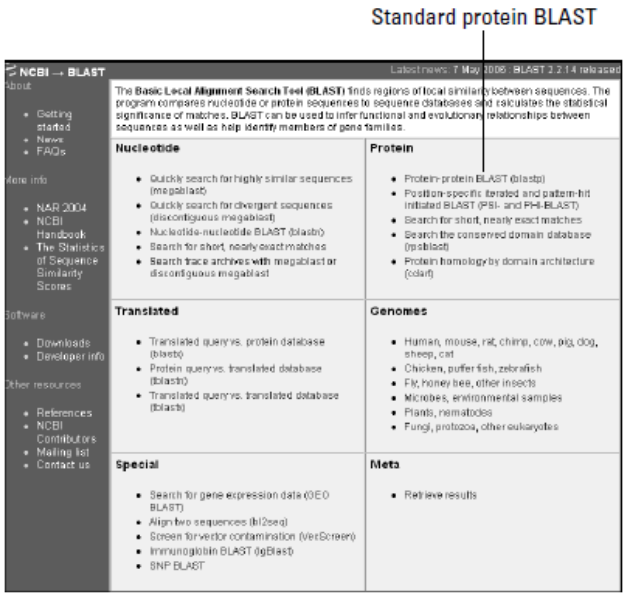
صفحه اصلی BLAST (که احتمالاً پر مخاطب ترین صفحه اینترنتی مرتبط با بیوانفورماتیک در جهان است) باز می شود (شکل ۲۴-۲).

۲. روی لینک Protein-Protein BLAST (blastp) در قسمت بالای سمت راست صفحه کلیک کنید.

صفحه جستجو باز می شود (شکل ۲۵-۲) حالا کافی است یک توالی پروتئین به فرم FASTA داشته باشید.

۳. فایلی که حاوی توالی پروتئین شما به فرم FASTA است را باز کنید.

این فایل همان فایلی است که با استفاده از اطلاعات موجود در بخش «یافتن توالی DNA مربوط به یک پروتئین خاص» در همین فصل در کامپیوتر خود ذخیره کرده اید.



شکل ۲۴-۲ : صفحه ورودی BLAST در NCBI

۴. با استفاده از گزینه های Paste , copy در «یکی از» توالی های پروتئینی را (همراه با خط تعریف آن) درون پنجره جستجوی BLAST وارد کنید.

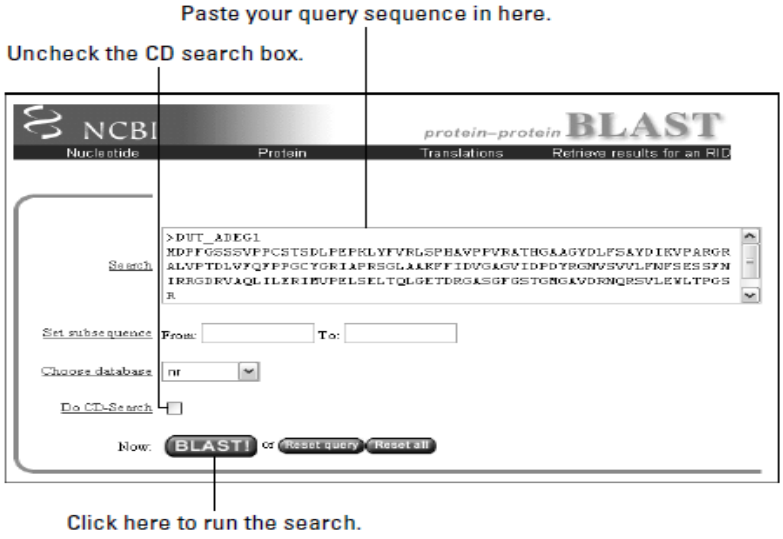
ما پروتئین DUT-ADEG1 را انتخاب می کنیم. اگر این توالی پروتئین را در اختیار ندارید، با انجام مراحل زیر می توانید به سرعت آن را در EXPASY پیدا کنید.

أ- وارد آدرس اینترنتی www.expasy.org/sprat/ شوید.

ب- کلید واژه DUT –ADEGI را در پنجره جستجوی بالای صفحه نوشته و روی دکمه search کلیک کنید.

ت- توالی پروتئینی را همراه با خط تعریف آن کپی کنید.

ث- پنجره را ببندید و توالی کپی شده را در پنجره جستجوی BLAST وارد کنید.



شکل ۲۵-۲: فرم جستجوی BLAST در NCBI

۵. مربع مقابل عبارت Do CD-search را از حالت انتخاب خارج کنید اما قسمت Choose Database را تغییر ندهید. nr [1] در پنجره database حاوی همه توالی های پروتئین شناخته شده است، بنابراین لازم نیست در این قسمت سازمان یا زیر مجموعه خاصی را انتخاب کنید.

۶. روی دکمه BLAST کلیک کنید.

اکنون شما اولین جستجوی خود را در BLAST انجام داده اید.

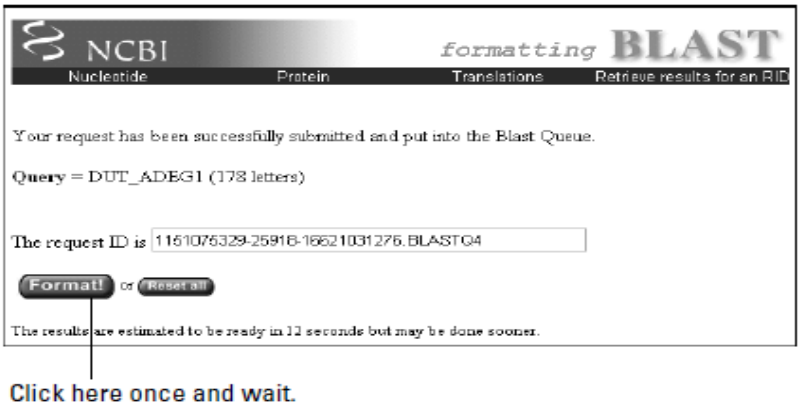
صفحه جدیدی با عنوان "formatting BLAST" باز می شود، که این مکان را به شما می دهد که تعدادی از گزینه ها را تغییر دهید (شکل ۲۶-۲). اما اکنون همه گزینه های موجود را دست نخورده باقی بگذارید.

۷. روی دکمه format کلیک کنید. نتایج جستجو ظرف چند ثانیه ظاهر می شوند.

۸. وقتی صفحه نتایج باز شد به طرف پایین صفحه حرکت کنید تا به فهرست بلندی از توالی ها برسید (شکل ۲۷-۲).

این فهرست شامل همه توالی هایی است که شباهت معنی داری با توالی پروتئین مورد نظر شمادارند و با شماره های نزولی رتبه بندی شده اند؟ اگر شما از DUT-ADEG1 یا هر توالی پروتئین دیگری که از بانک اطلاعاتی استخراج شده باشد استفاده کنید؟ احتمالاً مرتبط ترین پروتئین با این توالی اولین پروتئین در فهرست بدست آمده است. (مثلاً در شکل ۲۷-۲ مرتبط ترین پروتئین *fowl adenovirus dUTPase* است.) در فهرست شکل ۲۷-۲ بزرگترین عدد رتبه بندی ۳۶۱ است. شماره های رتبه بندی بستگی به طول شبیه ترین قطعات بین دو توالی و همچنین میزان شباهت آنها دارد البته توجه کنید که این شماره ۶ بر مبنای ۱۰۰٪ محاسبه نشده اند. (برای آشنایی بیشتر با شماره های رتبه بندی در BLAST فصل ۷ را ببینید.)

[1] nonredundant



شکل ۲۶-۲: فرم "format-while-you-wait" در بخش BLAST در NCBI

۹. روی یکی از شماره های رتبه بندی (ستون دوم از راست در شکل ۲۷-۲) کلیک کنید.

با این کار باز هم در طول صفحه پایین تر خواهید آمد و به چیزی شبیه شکل ۲۸-۲ خواهید رسید. اکنون شما می توانید میزان شباهت بین توالی مورد نظرتان و توالی پروتئین مربوط به شماره رتبه بندی که انتخاب کرده اید را بررسی کنید.

شکل ۲۷-۲: نتایج جستجو در BLAST: لیست بهترین توالی های مشابه

در مورد مثال ما، بهترین توالی (توالی با بزرگترین عدد رتبه بندی) دقیقاً شبیه توالی اولیه ما است. هرچه در فهرست توالی ها پایین تر بروید میزان شباهت توالی با توالی اولیه کاهش، عدد رتبه بندی کاهش و ارزش E[1] افزایش می یابد، ارزش E برآوردی از معنی داری آماری عدد رتبه بندی است. اگر ارزش E نزدیک ۱ باشد شما نمی توانید استنتاج قابل اطمینانی از میزان شباهت دو توالی داشته باشید. (در فصل ۷ توضیحات بیشتری در این مورد داده شده است).

شکل ۲۸-۲: نتایج جستجو در BLAST: تطابق نواحی از توالی ها

تطبیق چند توالی پروتئین با استفاده از برنامه ClustalW

پس از یافتن جفت پروتئین های مشابه کار دیگری که زیست شناسان معمولاً در زمینه بیوانفورماتیک پروتئین ها دوست دارند انجام دهند تطبیق دادن چند توالی پروتئینی با هم است. برای این کار تعداد زیادی توالی پروتئینی در کنار هم قرار داده شده و مقایسه می شوند.

با تطبیق دادن چند توالی پروتئینی می توانید:

- ✓ محل هایی از توالی که آمینو اسیدهای آنجا نقش مهمی در یکپارچگی ساختاری یا عملکرد یک پروتئین خاص دارند را مشخص کنید.
- ✓ یک توالی خاص را مشخصه یک خانواده پروتئینی قرار دهید.

[1] E-value

✓ توالی ها را دسته بندی کنید و درخت های تکاملی برای آنها رسم کنید.

جزئیات بیشتر در مورد چگونگی تطبیق دادن چند توالی پروتئینی در فصل ۹ بحث خواهد شد. در این بخش هدف ما فقط اینست که به شما نشان دهیم که تطبیق دادن چند توالی کار آسانی است، بخصوص زمانیکه از یک سرور اینترنت خوب مثل PIR (Protein Information resorce) استفاده می کنید. این سرور در دانشگاه Georgetown واشنگتن قرار دارد.

PIR[1] در حقیقت از اولین مجموعه توالی های پروتئینی با نام Atlas of protein sequences (که توسط پروفسور M.Dayhoff در اواخر دهه ۱۹۷۰ ساخته شده است) منشاء گرفته است. سایت PIR چند ابزار و بانک اطلاعاتی مفید برای آنالیز پروتئین ها در اختیار شما قرار می دهد و ما پیشنهاد می کنیم که از آن استفاده کنید. در بین ابزارهای موجود در این سایت، یک سرویس برای تطبیق چند پروتئین وجود دارد. این سرویس از برنامه clustalW استاندارد استفاده می کند. برای شروع کار با clustalW مراحل زیر را انجام دهید:

۱. آدرس اینترنتی pir.georgetown.edu را جستجو کنید.
صفحه اصلی سایت PIR ظاهر می شود (شکل ۲۹-۲).
۲. منوی مربوط به Search/Analysis را باز کرده گزینه Multiple Alignment را انتخاب کنید. یک فرم ظاهر می شود (شکل ۳۰-۲). اکنون شما نیاز به چند توالی پروتئین با فرمت FASTA دارید.
۳. فایل مربوط به توالی dUTPase را که با فرمت FASTA در کامپیوتر خود ذخیره کرده اید (در بخش های قبلی این فصل این کار را انجام دادید)، باز کنید.

اگر توالی dUTPase را در کامپیوتر خود ندارید، می توانید از طریق سایت Swiss-Prot (www.expasy.org/sprat/) چند توالی مرتبط با dUTPase (مثلاً DUT-AQUAE ، DUT-BRAJA ، DUT-BPT5 ، DUT-CANAL ، DUT-BUCAI) را به روشی که در بخش "ExPASy" این فصل توضیح داده شد پیدا کنید.

۴. پنج تا از توالی های بدست آمده را در پنجره موجود در فرم (شکل ۳۰-۲) وارد کنید.
توجه کنید که در هر مورد «خط تعریف» و توالی کامل آمینو اسیدی پروتئین ها را وارد کنید.
۵. روی دکمه submit کلیک کنید.

پس از چند لحظه شما می توانید اولین مقایسه بین توالی های چند پروتئین و یک شمای درخت مانند که طول توالی ها را با هم مقایسه می کند را ببینید (شکل ۳۱-۲).

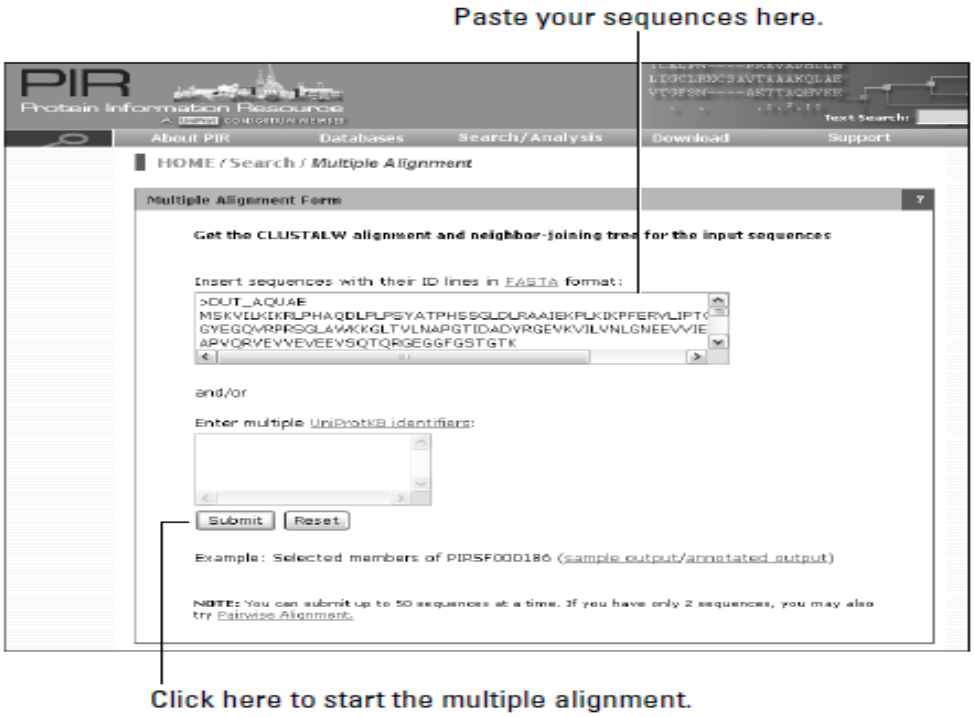
با تطبیق دادن چند توالی با هم:

✓ قسمت هایی از توالی این پروتئین ها که شباهت بیشتری با هم دارند، مشخص می شود. موقعیت هایی از توالی پروتئین ها که ۱۰۰٪ شبیه هم هستند (که با علامت * مشخص شده اند) و موقعیت هایی که توسط آمینو اسیدهای «مشابه از نظر شیمیایی» اشغال شده اند بیشتر تمایل دارند که در کنار هم واقع شوند. این موقعیت ها احتمالاً نواحی را بوجود می آورند که مستقیماً در شکل گیری محل فعال آنزیم مشارکت دارند یا واکنش بیوشیمیایی پیرو فسفاتاز را انجام می دهند.
✓ انواع گوناگونی نواحی عملکردی۲ (قطعات کوچکی از توالی که با یک عملکرد بیوشیمیایی خاص مرتبطند) مشخص می شوند.

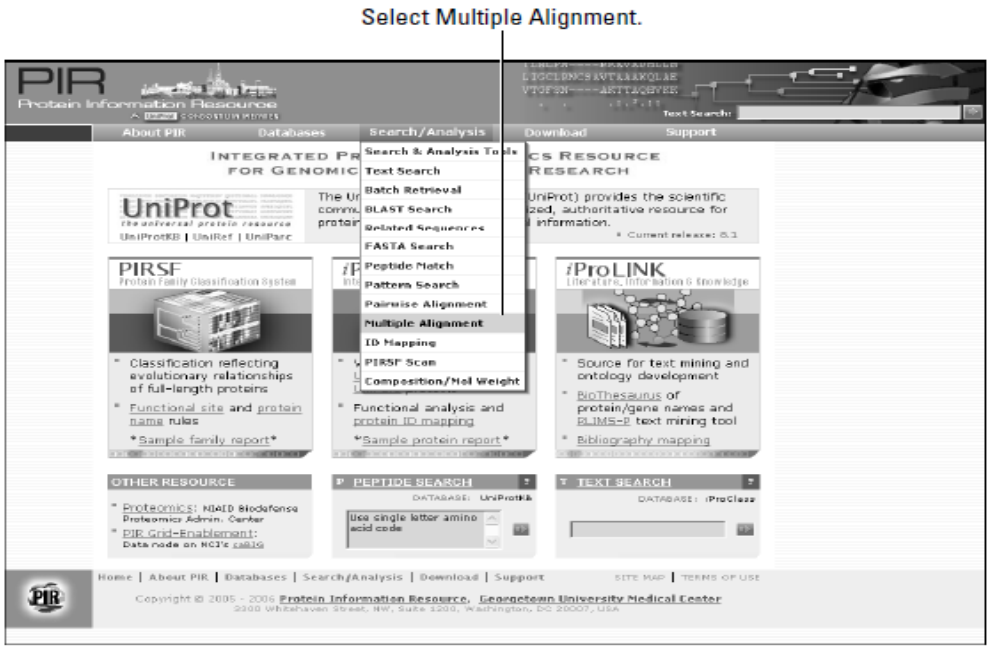
[1] protein information resource
[2] Functional signatures

✓ می توان با استفاده از شجره های تکاملی، روابط تکاملی بین پروتئین ها را استنتاج کرد.

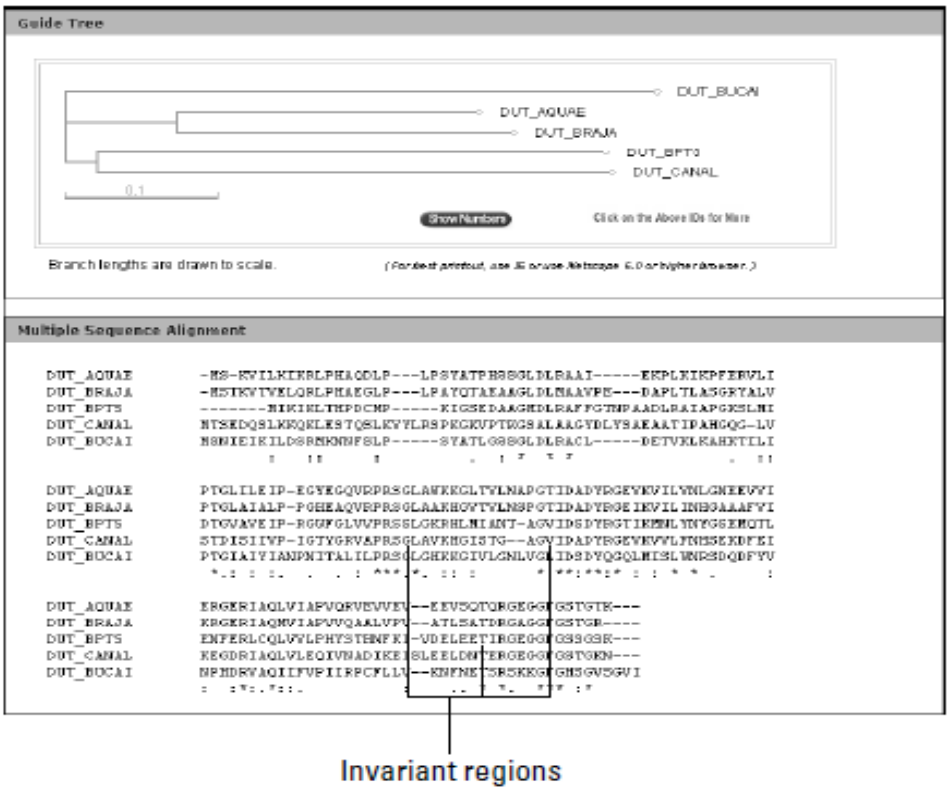
در فصل های ۹ ، ۱۱ و ۱۳ در مورد این کاربردهای کلیدی بیوانفورماتیک بیشتر بحث خواهد شد.



شکل ۲۹-۲: یک مفید مرکز آنالیز پروتئین دیگر: PIR



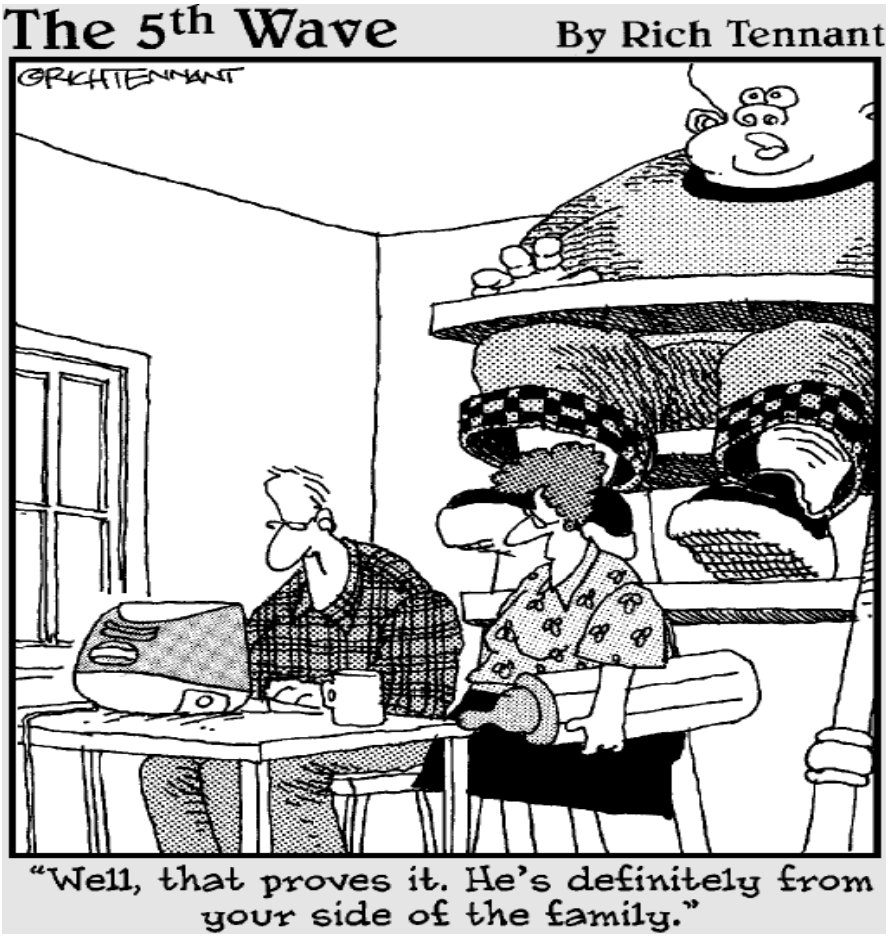
شکل ۳۰-۲: پر کردن فرم مربوط به تطبیق چند توالی در PIR



شکل ۳۱-۲: تطابق توالی های پنج dUTPase باکتریایی و درخت ایجاد شده در اثر تشابه جفت به جفت آمینواسید ها

بخش دوم

راهنمایی های ضروری در بیوانفورماتیک



ترجمه: سیاوش ایروانی

ویرایش: دکتر حسن کربکندی، سجاد عباسی

در این بخش می خوانیم ...

بیوانفورماتیک، جستجو و تفسیر اطلاعات زیست شناسی بصورت آنلاین است. در این قسمت چگونگی استفاده از اطلاعات موجود در بانک های اطلاعاتی بیوانفورماتیک نشان داده خواهد شد. همچنین اطلاعات و نکات ریز بسیاری بصورت تازه و آماده برای استفاده وجود دارد (نه فقط میوه های حاضر و آماده برای چیدن!). از طرفی طریقه یافتن اطلاعات خاص از میان انبوهی از اطلاعات را فرا خواهید گرفت.

این یک سفر دل انگیز و جذاب میان انبوهی از اطلاعات و دانش بشری است و نقطه شروعی اجباری برای هر پروژه تحقیقاتی. اطلاعات مورد نظر در یخچالند، کافی است آنها را از یخچال در آورده و شروع به پختن نمایید.

بعلاوه در این قسمت دستوالعمل های ساده برای کار با توالی های DNA و پروتئین به منظور پیش بینی و دریافت اطلاعات اساسی و پایه، در اختیارتان قرار می گیرد.

فصل سوم

استفاده از بانک های اطلاعاتی توالی های نوکلئوتیدی

در این فصل می خوانیم :

- ✓ دریافت اطلاعات جدید درباره ساختار ژنها و ژنوم ها
- ✓ نحوه استفاده و درک صحیح از GenBank
- ✓ دریافت اطلاعاتی پیرامون ژنهای اختصاصی
- ✓ کار با ژنوم های میکروبی کامل
- ✓ نگاهی به اطلاعات مربوط به ژنوم انسان و دیگر موجودات در کامپیوتر

.....

" راز موفقیت، یافتن آنچیزی است که دیگران از آن بی خبرند. "

— Aristotle Onassis (190 –1975)

بانک های اطلاعاتی توالی ها، بهترین ابزار برای گشودن پنجره ای به گذشته است. آنالیز این توالی ها، حتی اگر مربوط به ۲۵ سال پیش باشند، امکان پاسخگویی به بسیاری از پرسشهای بیولوژی نوین را برقرار می سازند، به خصوص زمانی که نیاز به مرور همه فناوری های موجود احساس شود. بنابراین امکان اتصال بیولوژی مولکولی گذشته وامروز برقرار می گردد. اولین بانک های اطلاعاتی، به عنوان یک موزه، تمامی توالی ها را دقیقا مشابه زمانی که برای اولین بار توسط محققین خود شناسایی، تفسیر و انتشار پیدا کردند، در خود جای می دهند.

این دورنمای تاریخی (کپسول زمان) در بانک اطلاعاتی GenBank بطور کاملا شکیل جلوه می نماید. GenBank بهترین مجموعه از توالی های نوکلئوتیدی است که به شکل کنسرسیوم مشترک میان مرکز بین المللی اطلاعات بیوتکنولوژی آمریکا[1] (NCBI)، کتابخانه بیولوژی مولکولی اروپا[2] (EMBL) و بانک اطلاعاتی DNA ژاپن[3] (DDBJ) قرار دارد. در این فصل، نحوه استفاده از GenBank و کشف ورودی های موجود در آن نشان داده خواهد شد.

بانک های اطلاعاتی ابزاری مناسب برای جستجوی منابع مرتبط با یک توالی خاص می باشند ولی دسترسی به این اطلاعات زمانی آسان است که حوزه وسیعی از ژن و یا ساختار آن را شامل نشود. به همین دلیل نسل دوم بانک های اطلاعاتی توالی های نوکلئوتیدی با رویکرد بیشتر به ژن و قابلیت دسترسی سریع به همه اطلاعات توالی های ژنی شکل گرفتند. از طرف

[1] the U.S. National Center for Biotechnology Information
[2] the European Molecular Biology Laboratory
[3] the DNA Data Bank of Japan

دیگر چگونگی رمز گشایی نسخه های قدیم تر GenBank، و نیز استفاده از NCBI's Entrez Gene (یک مثال خوب از بانک های اطلاعاتی با محوریت ژن) را به شما نشان می دهیم. این روز ها، توالی های نوکلئوتیدی در گستره ژنوم یا کروموزوم به راحتی قابل شناسایی است – حداقل برای میکرواورگانیسم ها. اکنون ما نه تنها اطلاعاتی درباره توالی ژنهای منفرد بلکه اطلاعاتی پیرامون جایگاهای وابسته (خویشاوندی)، جهت گیری رشته DNA ، وجود یا عدم وجود کارکردهای بیوشمیایی در یک ارگانیسم را در اختیار داریم. به منظور بهره برداری از فواید این اطلاعات بی نظیر، محققین مجبور به ساخت سیستم های پردازش کننده اطلاعات توالی ها با محوریت ژن هستند، تا با انتخاب ابزار مناسب امکان برقراری اتصال با مجموعه ای از توالی های خاص وجود داشته باشد. به عنوان پاداش ، بعضی از منابع بی نظیر با محوریت ژن را که مرتبط با ویروس ها، باکتری ها و یا انسان می باشد را در اختیارتان قرار می دهد.

با این وجود قبل از کاوش در این سیستمهای پردازشگر اطلاعات، ابتدا باید قسمت بعدی را نیز مطالعه کنید. ما در آن قسمتها، به سرعت و بطور خلاصه، یافته ها و اطلاعات ژن و ژنوم را برای شما بازگو کرده و واژگان و اصطلاحاتی راکه به هنگام کار با GenBank به آن نیاز پیدا خواهید کرد، معرفی می کنیم. در صورتیکه یک زیست شناس با تجربه هستید، می توانید از این قسمت بطور کامل عبور کرده و آن را نادیده بگیرید و یا اینکه اصطلاحات را به دقت خوانده و رفع اشکال نمایید! پس معتل چه هستید شروع کنید، ما نیز در این راه به شما کمک می کنیم. اگر تمایل به یادگیری بیشتر در زمینه ژنتیک را دارید، به شما کتاب ژنتیک به زبان ساده، نوشته Tara Rodden Robinson، را پیشنهاد می کنیم.

مطالعه در زمینه ژنها و ژنوم

این درست است که توالی های نوکلئوتیدی همگانی هستند ولی ساختار آنها و نحوه ی کدگذاری آنها در بین پروکاریوت ها (موجوداتی فاقد کروموزوم حقیقی) و یوکاریت ها (موجوداتی دارای کروموزم حقیقی) بسیار متفاوت است. شما باید بر ساختار ژنوم و ژنهای تک تک آنها مسلط باشید، تا به درک درستی از GenBank برسید. برای این منظور نیاز به شناسایی ژن به شکل کلیه قطعات ژنوم مربوطه که شامل اطلاعات ضروری همه توالی های نوکلئوتیدی مورد نیاز برای یک بیان کامل و موفق(یا همان تولید پروتئین و یا RNA) می باشد. این شناسایی بخش های کدکننده و نیز بخش های تنظیم کننده یک ژن را شامل می شود.

پروکاریوت ها: موجودات کوچک ، ژن های ساده

سه دسته عمده از موجودات زنده شامل پروکاریوت ها (به طور عمده باکتری ها)، Archaea (موجوداتی شبه باکتری که در شرایط سخت قابلیت زندگی دارند) و یوکاریوتها . یوکاریوت ها شامل انواعی از مخمرهای میکروسکوپی تا انسان ، حیونات و گیاهان می باشند.

به منظور پیشبرد اهداف بیوانفورماتیک، پروکاریوت ها و جلبک ها را شبیه بهم در نظر گرفتیم البته با بعضی استثنا که فراتر از حوزه این کتاب است. آنها دارای خصوصیات مشترک زیر هستند:

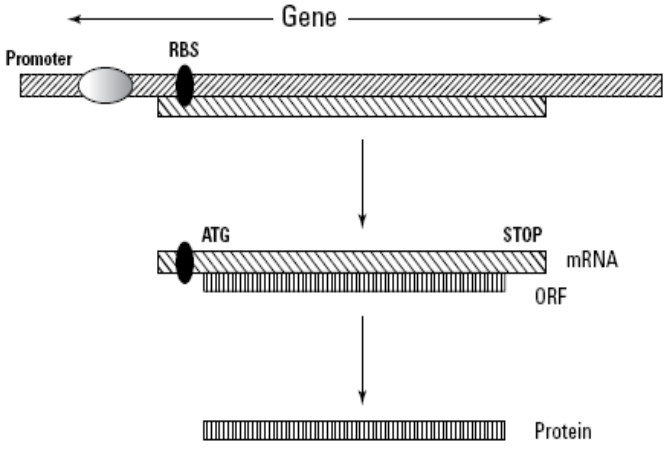
- ✓ موجوداتی میکروسکوپی هستند.
- ✓ ژنوم آنها یک مولکول DNA آنها حلقوی منفرد است.
- ✓ ژنوم آنها از چند میلیون جفت باز تشکیل شده است (۰/۶–۸) .
- ✓ دانسیته ژن (نسبت تعداد ژن بر جفت بازهای موجود در ژنوم) در حدود ۱ ژن بر ۱۰۰۰ جفت باز می باشد.

- ✓ ژنوم مختصر و جزئی دارند، و بخش های بی استفاده ژنوم آنها ناچیز می باشد(۷۰٪ آن به منظور کدشدن برای تولید پروتئین بکار می رود).
- ✓ ژنهای آنها هیچگونه همپوشانی ندارد.
- ✓ ژنهای آنها درست بعد از ناحیه ای تنظیمی به نام پروموتور[1] به mRNA [2] نسخه برداری می شود.
- ✓ mRNA با توالی ژنوم معادل است. به عبارت دیگر قطعات ژنی به صورت یکپارچه قرار گرفته و توسط مناطقی که کد نمی شوند (به اصطلاح اینترون)، از یکدیگر جدا نشده اند.
- ✓ توالی های آمینو اسیدی، در نتیجه ترجمه بزرگترین توالی های باز خواندن (ORF) موجود در mRNA ایجاد می شوند.

شکل (۱-۳) با مثال، ارتباط خطی و مستقیم بین توالی ژنوم باکتریایی، mRNA نسخه برداری شده، ORF و پروتئین ساخته شده را نشان می دهد.

برای مشاهده سریع شماتیک یک ژنوم معمول باکتری ، به شکل ۱۰-۱ در فصل ۱ سری بزنید. پس از یک سیگنال خاص از محل اتصال ریبوزوم[3] (RBS در شکل ۱-۳)، توالی mRNA به پروتئین مورد نظر ترجمه می گردد. ریبوزوم اصلی ترین قسمت دستگاه ترجمه در سلول می باشد. از این رو یک بانک اطلاعاتی برای توصیف یک توالی کدکننده پروکاریوتی باید به سه مشخصه زیر بپردازد:

۱- مختصات بعضی از پروموتور ها
۲- مختصات RBS
۳- مختصات و حدومرز ORF



شکل ۱-۳: ارتباط بین ژن، mRNA و توالی پروتئین پروکاریوت ها

همه آنچه گفته شد توصیفی از ساختمان و معماری ساده ژنهای پروکاریوتی می باشد. همه ژنها به پروتئین تبدیل نمی شوند، در بعضی موارد، همان مولکول RNA نسخه برداری شده عملکرد مورد نظر را انجام می دهد، مثلا RNA ناقل (tRNA) و RNA ریبوزومی (rRNA) و بعضی مولکولهای RNA دیگر، که شاید بهتر باشد از آنها نامی برده نشود.

---

[1] promoter
[2] messsenger
[3] Ribosome Binding Site

یوکاریوت ها: موجودات بزرگ تر، ژن های پیچیده تر

یوکاریوت ها گستره وسیعی از موجودات زنده را شامل می شوند، از مخمر های میکروسکوپی گرفته تا قارچها و موجودات بزرگتری چون گیاهان، حیوانات و انسان ها. یوکاریوت ها همچنین دارای خصوصیات اساسی مشترک زیادی هستند. هنگامی که یوکاریوت ها تحت آنالیز های بیوانفورماتیکی و ژنومیکی قرار می گیرند، این خصوصیات موجب ایجاد انگیزه ای بیشتر برای بحث پیرامون آنها می شود:

- ✓ ژنوم آنها شامل قطعات خطی از DNA به نام کروموزم است که طول آنها بیش از یکصد میلیون جفت باز است.
- ✓ اندازه ژنوم آنها ۱۰-۶۷۰۰۰۰ میلیون جفت باز است، خصوصا گیاهان و حیوانات و آمیب (!) که اندازه ژنوم آنها بسیار بزرگتر از پروکاریوت ها است.
- ✓ دانسیته ژنی آنها از پروکاریوت ها بسیار کمتر است (مثلا در انسان یک ژن بر ۱۰۰۰۰۰ جفت باز).
- ✓ ژنوم آنها شامل قسمت های بدون استفاده زیادی است (کمتر از ۵٪ از ژنوم انسانی به پروتئین کد می شود) .
- ✓ ژن های موجود در رشته های مقابل هم DNA ممکن است همپوشانی پیدا کنند ، اگر چه این اتفاق به ندرت رخ می دهد.
- ✓ نسخه برداری درست بعد از یک ناحیه تنظیمی به نام پروموتور شروع می گردد ولی وجود عوامل موجود در توالی که در جایی دورتر قرار گرفته اند در این پروسه نقش موثری را ایفا می کنند (عوامل موثر در فرایند نسخه برداری).
- ✓ توالی های ژنی با mRNA و توالی های پروتئین معادل نیستند. فقط بعضی از قسمت ها به نام اگزون، در mRNA باقی می مانند و محصول نهایی از روی آن کدگذاری می شود.
- ✓ ژنها اغلب به بیش از یک نوع mRNA و پروتئین تبدیل می گردند.

ژن های یوکاریوت های عالی تر (حیوانات)، از چندین میلیون جفت باز تشکیل شده است. برای مثال طول humen dystrophin gene، که جهش در آن باعث ایجاد یک بیماری خطرناک می شود،۲/۲ میلیون جفت باز است. ارتباط بین توالی های DNA در ژن، نسخه ابتدائی آن، گونه های متفاوت از mRNA بالغ و توالی پروتئین نهایی، می تواند بسیار پیچیده باشد. به همین دلیل است که درک صحیح یک ورودی بانک اطلاعاتی که همه وقایع بالا را توصیف می کند، نیاز به کمی کار دارد. به علاوه، بسیاری از این ورودی ها با بخشی از توالی های ژن و یا با اجزای مرتبط با ژن (مانند نواحی پروموتور، mRNA و قطعات ژنوم) در ارتباطند. در ادامه این فصل، با مثال های متنوعی برخورد خواهید کرد که با کمی مطالعه آنها، به تنهایی قادر به کشف بسیاری از ورودی های پیچیده GenBank خواهید بود.

نحوه استفاده از **GenBank**

این واقعیت که پروکاریوت ها با ژنهای با محدوده خاصی از اندازه سر و کار دارند، و همچنین ارتباط خطی و ساده ای بین توالی DNA ژن، mRNA، ORFs و توالی پروتئین نهایی وجود دارد، موجب می گردد تا تمامی اطلاعات مرتبط با آنها به آسانی در بانک های اطلاعاتی ذخیره وثبت گردند. به همین دلیل مطالعه و درک اطلاعات مرتبط با ژن های باکتریایی اغلب آسان است. در این قسمت، ورودی پروتئینdUTPase باکتری *Escherichia coli* (با شماره شناسه X01714 در بانک اطلاعاتی GenBank) را به صورت مرحله به مرحله بررسی خواهیم کرد.

درک صحیح از ورودی های GenBank پروکاریوتی

برای پیمودن راه ابتدا یک ورودی GenBank را مطابق دستوالعمل زیر بررسی می کنیم:

۱. به آدرس اینترنتی www.ncbi.nlm.nih.gov/entrez/ رجوع کنید.
صفحه اصلی سایت NCBI PubMed ظاهر می گردد.

۲. از منوی Search، گزینه Nucleotide را انتخاب کنید (مطابق شکل ۲-۳)

۳. در قسمت پنجره For، عبارت GenBank ID X01714 را تایپ کرده و سپس بر روی Go کلیک نمائید.
صفحه ظاهر شده مختصری از اطلاعات ورودی توالی نوکلئوتیدی GenBank ID X01714 را نشان می دهد (با زیر خط آبی).

۴. بر روی لینک X01714 کلیک نمایید(به عنوان یک راه فرعی، می توانید قسمت Display را از حالت Summary به GenBank تغییر دهید).
ورودی GenBank X01714، همانطور که در شکل ۳-۳ نشان داده شده است، ظاهر می گردد. معلمین این فرمت را، فرمت flat-file بانک اطلاعاتی GenBank می نامند، که فاقد هر گونه فهرست راهنما، اشاره گر و یا اطلاعات کمکی،که در بانک های اطلاعاتی حرفه ای و پیچیده مرسوم هستند، است. در حقیقت آنچه اینجا مشاهده می کنید آنقدرها هم flat نیست، چون تعدادی لینک در آن وجود دارد.

۵. گزینه text را از منوی sent to انتخاب کنید تا ورودی X01714 با فرمت text یا flat-file نمایان گردد.

۶. با انتخاب گزینه save as از منوی file، متن مورد نظر را در کامپیوتر خود ذخیره نمایید.

شکل ۲-۳: جستجوی توالی نوکلئوتیدی X01714 در NCBI PubMed

خواندن سربرگ یک ورودی GenBank پروکاریوتی

متن حاوی اطلاعات ورودی مورد نظر که در مرحله قبلی ذخیره شده بود، به بخش هایی تقسیم می گردد، و هر بخش توسط یک کلید واژه معرفی می گردد. کلید واژه های معمول عبارتند از LOCUS، DEFINITION، ACCESSION، VERSION، KEYWORDS و SOURCE (شکل۳-۳).

در این قسمت مرحله به مرحله کلید واژه ها را تعریف می کنیم :

- ✓ **LOCUS:** اطلاعاتی راجع به نام مکان مورد نظر روی DNA (یک نام قراردادی، که در این مثال ECDUT می باشد) ، سایز و اندازه توالی نوکلئوتیدی بر حسب جفت باز، ماهیت مولکول (مولکول DNA) و وضعیت قرارگیری آن (خطی یا حلقوی بودن، که در این مورد مولکول ما خطی است) به ما می دهد.
- ✓ **DEFINITION:** تعریف کوتاهی از ژنی که توالی مورد نظر ما درون آن قرار گرفته است، ارائه می دهد. در اینجا ژن مورد نظر dUTPase *E. coli* می باشد.
- ✓ **ACCESSION:** حاوی لیستی از شماره های دسترسی[1] است. این شماره ها شناساگرهای بی نظیری هستند که در یک بانک اطلاعاتی خاص و در بین چند بانک اطلاعاتی(از جمله بانک های اطلاعاتی توالی های پروتئین) قابل استفاده هستند. در واقع با استفاده از این شماره می توانید به سرعت یک ورودی خاص را پیدا کنید.دراینجا شماره دسترسی X01714 می باشد.
- ✓ **VERSION:** در این قسمت اطلاعاتی راجع به مترادف ها و شماره شناسایی گذشته در اختیارتان قرار خواهد گرفت.
- ✓ **KEYWORDS:** لیستی از واژه هایی که به خوبی معرف ورودی مورد نظر هستند را معرفی می نماید. می توانید از این واژه ها برای جستجو در مورد یک بانک اطلاعاتی خاص استفاده کنید.. برای استفاده بیشتر از کلمات کلیدی و پی بردن به نقشی که در جستجوی داده ها ایفا می نمایند به فصل ۲ رجوع کنید، در آنجا به تفصیل در مورد استفاده از PubMed و محدودیت ها ی آن بحث شده است.
- ✓ **SOURCE:** در این قسمت نام عمومی ارگانیسمی که توالی مورد نظر به آن تعلق دارد، بیان گردیده است.
- ✓ **ORGANISM:** اطلاعات جامعی را درمورد ارگانیسم مورد نظر بیان می کند.به علاوه اطلاعات مربوط به دسته بندی و تاکسونومی[2] موجود را در اختیارتان می گذارد.

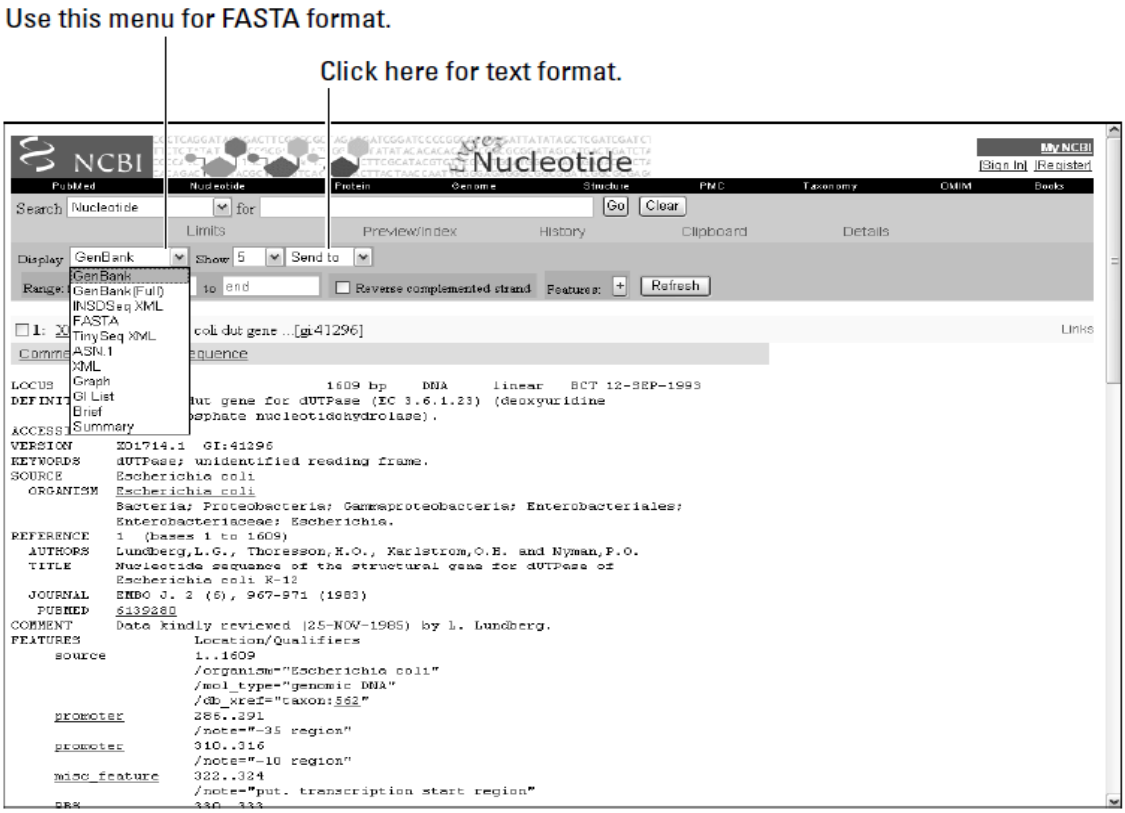
توجه داشته باشید که قسمت ORGANISM نسبت به حاشیه سمت چپ صفحه فاصله کمی بیشتر از دیگر کلمات کلیدی دارد. این یک خطا نیست، متخصصان کامپیوتر به آن تورفتگی[3] می گویند. اینکه یک کلید واژه در موقعیتی جلوتر یا عقب تر از کلید واژه دیگر قرار می گیرد، نشاندهنده آن است که یک کلید واژه زیرمجموعه ای از کلید واژه دیگر است. حال آنکه قرار گرفتن کلید واژه ها از مکانی مشابه نشاندهنده یکسانی اهمیت آنهاست. کلید واژه هایی که در منتهاعلیه سمت چپ قرار گرفته اند، مثل ۶ کلید واژه ابتدایی که توضیح داده شدند، نشاندهنده بخش های مستقل ورودی موجود در GenBank هستند. اما بخش ORGANISM زیر مجموعه ای از بخش SOURCE تلقی می شود (بخش REFERENCE هم به همین صورت است).

- ✓ **REFERENCE:** در این قسمت توالی های شناسایی شده از لحاظ اعتبار بررسی می گردند (قسمت های مختلف توالی ها می توانند نویسندگان متفاوتی داشته باشند). بخش REFERENCE حاوی قسمت های گوناگونی است که

[1] accession numbers
[2] taxonomic classification
[3] indentation

بوسیله کلمات کلیدی مشخص از یکدیگر جدا شده اند، مانند: AUTHORS ، TITLE، JOURNAL و PUBMED.

- ✓ **COMMENT:** این بخش، بخش انتهایی قسمت بالایی ورودی GenBank X01714 است، و حاوی متن هایی با فرمت آزاد شامل تقدیر و تشکر[1] و همچنین اطلاعاتی است که در بخش های قبلی ذکر نشده اند.



شکل ۳-۳: ورودی X01714 از بانک اطلاعاتی GenBank

قسمت بعدی ورودی (قسمت میانی)، جدول مشخصه ها[2] است (شکل ۴-۳). این بخش به طور صریح به بررسی نواحی ژنی و خصوصیات بیولوژیکی مرتبط با آن ها که در توالی نوکلئوتیدی قابل شناسایی است، می پردازد. کلید واژه کلی برای این بخش FEATURES می باشد و یک زیر مجموعه بزرگ از واژگان کلیدی را شامل می شود که ما آنها را مطابق لیست زیر بیان می کنیم.

- ✓ **source:** این بخش در حقیقت منشأ یک بخش خاص از توالی را مشخص می کند. این بخش به منظور شناسایی وکتور[3]های کلون کننده در توالی میزبان کاربرد دارد. در مورد X01714، منشأ توالی تماما از DNA ژنومی *E. coli* می باشد.
- ✓ **promotor:** مختصات دقیق عوامل آغازگر را به ما نشان می دهد. در X01714، منطقه a-35، از مکان ۲۸۶ تا ۲۹۱ در توالی نوکلئوتیدی را مشخص می نماید، که به صورت 291.. 28 نمایش داده می شود. علاوه بر این مختصات یک عامل آغازگر دیگر را نیز با نام ناحیه a-10 درمکان ۳۱۶-۳۱۰ مشخص کرده است.

[1] acknowledgments
[2] features table
[3] vector

- ✓ **misc feature:** به معنی شاخصه های متفرقه[1] است، در این بخش مکان های مشهوری که نسخه برداری از آنها شروع می شود، نشان داده شده اند. مثلا در مورد X01714، این مکان از ۳۲۲ تا ۳۲۴ می باشد.
- ✓ **RBS** (جایگاه اتصال ریبوزوم): نشاندهنده مکان آخرین عنصر بالادستی[2] در رشته DNA می باشد. مثلا در مورد X01714، مکان ۳۳۰ تا ۳۳۳ را به خود اختصاص می دهد.
- ✓ **CDS** (قطعه کدکننده): نشاندهنده جایگاه پیچیده ای است که *قالب باز خواندن (ORF)* را نشان می دهد.
  - اولین خط مختصات ORF را، از نقطه شروع آن که ATG می باشد تا آخرین نوکلئوتید از اولین کدون پایان که TAA می باشد، نشان می دهد. برای مثال در مورد X01714، از ۳۴۳ تا ۷۹۸ را به خود اختصاص می دهد.اگر به خاطر نمی آورید که کدون چیست به فصل ۱ رجوع کنید.
  - هریک از خطوط پایین تر، که همگی در یک سطح هستند، نام یک محصول پروتئینی را عرضه می کنند، و نشاندهنده قالب خواندن (در اینجا ۳۴۳ اولین باز از اولین کدون است) و کد ژنتیکی به کار رفته، و همچنین شماره شناسایی توالی پروتئینی است.
  - translation، کلید واژه نهایی از بخش CDS می باشد که یک توالی آمینواسیدی مجازی از ناحیه کد کننده ارائه می دهد. این توالی، در حقیقت ترجمه کامپیوتری است که با استفاده از تمامی مختصات، قالب خواندن و کدهای ژنتیکی ارائه شده است.

یکی دیگر از شاخصه های متفرقه بعد از CDS، بخشی است که به ساختار ساقه–حلقه[3] و تکرار ها اشاره می کند. این ها همان عوامل بالقوه تنظیم کننده در ژن dUTPase می باشند.

بخش FEATURES در مورد ورودی X01714 ، یک نمونه ساده و تمام وکمال از توالی نوکلئوتیدی باکتریایی است، که حول یک ژن شناخته شده مرکزیت دارد. با این وجود، این ورودی های سرراست، یک پیچیدگی نیز دارد: یک قالب خواندن اضافی و مشهور. ما این ورودی را به این خاطر انتخاب کردیم که به ژن dUTPase علاقمند بودیم. با این حال، این ورودی حاوی یک ژن خاص و مشهور است که ویژگی آن یک RBS اضافی و یک بخش CDS ثانویه است.

ورودی های حاوی بیش از یک ژن هم در GenBank معمول هستند.

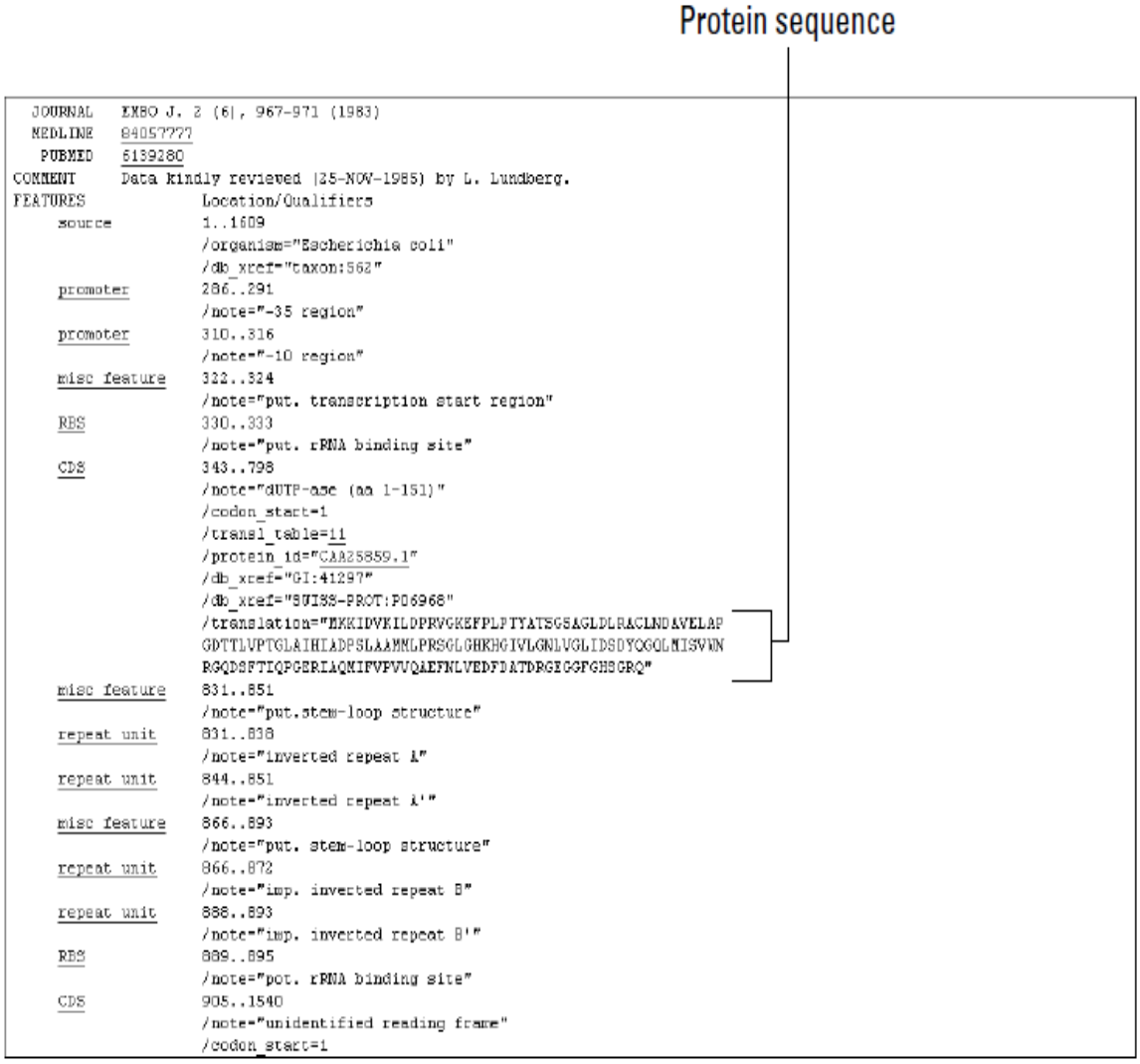
استفاده از قسمت Sequence یک ورودی پروکاریوتی

آخرین قسمت ار ورودی X01714 در GenBank، بخش توالی نوکلئوتیدی است. این قسمت با واژه کلیدی ORIGIN شروع شده و با نشان دوخط مورب (//) به اتمام می رسد. هر خط از بخش توالی نوکلئوتیدی بوسیله شماره مکان اولین نوکلئوتید آن خط شروع می گردد، و شامل ۶۰ نوکلئوتید می باشد.

به دلیل اینکه شماره ها و نوکلئوتید بطور همزمان در بخش توالی یک ورودی GenBank درج شده اند، شما قادر به استفاده از این توالی ها به عنوان یک داده مناسب برای اغلب سرورهای آنالیز کننده توالی، نخواهید بود.برای این کار، ابتدا باید توالی با فرمت FASTA را مطابق مراحل زیر بسازید:

---

[1] miscellaneous feature
[2] upstream
[3] stem-loop



شکل ۴-۳: ورودی Genbank X01714 : جدول شاخصه ها

۱. بر روی صفحه مربوط به ورودی GenBank، از منوی Display گزینه FASTA را انتخاب کنید(شکل ۳-۳). منوی Display در قسمت چپ و بالای صفحه قرار دارد.

۲. گزینه Text را از منوی Send to انتخاب کنید. اکنون فرم توالی نوکلئوتیدی برای آنالیز توسط اغلب برنامه های آنالیز توالی مناسب است.

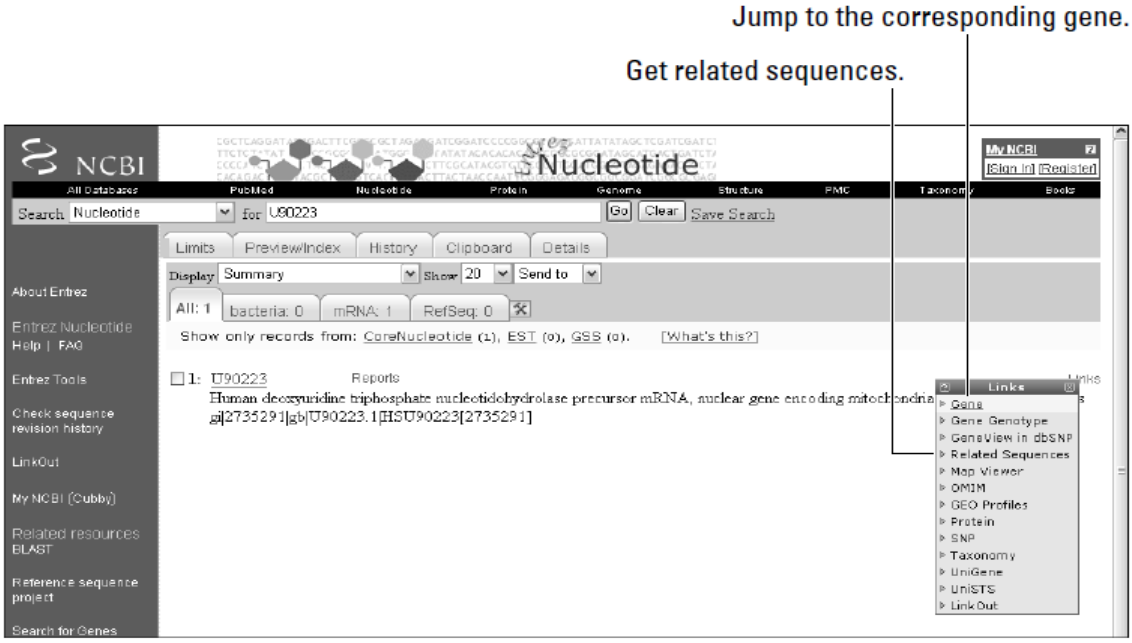
۳. این اطلاعات را انتخاب گزینه File از منوی اصلی و سپس گزینه Save as و یا از طریق کپی و paste کردن آن در یک فایل Word ذخیره نمایید.

درک مفهوم ورودی یک mRNA یوکاریوتی در GenBank

ما اکنون با ژن dUTPase کار خود را ادامه می دهیم چون بسیار جامع و فراگیر بوده ودر پروکاریوت ها و یوکاریوت ها وجود دارد، و همچنین در انسان نیز یافت می شود. نگاهی به ورودی های مربوط به نسخه انسانی dUTPase در

GenBank، به وضوح سطح بالاتری از پیچیدگی را در ژن های موجودات عالی یوکاریوت نسبت به پروکاریوت ها نشان می دهد. ما در اینجا با نوع ساده ای از آن کار را شروع می کنیم، ورودی U90223 در GenBank:

۱. به آدرس اینترنتی www.ncbi.nlm.nih.gov/entrez/ مراجعه کنید. صفحه اصلی NCBI PubMed ظاهر می گردد.
۲. از روی منوی Search گزینه Nucleotide را انتخاب کنید.
۳. در پنجره For، عبارت GenBank ID U90223 را تایپ کنید و سپس بر روی گزینه Go کلیک نمایید. صفحه نتایج ظاهر می گردد و در آن تحت عنوان ID U90223 تعریف کوتاه و مختصری از ورودی توالی نوکلئوتیدی ظاهر می گردد (شکل ۵-۳).
۴. بر روی لینک U90223 کلیک کنید(به عنوان یک راه فرعی، می توانید حالت Display را از Summary به GenBank تبدیل کنید). اکنون می توانید اولین ورودی GenBank انسانی را ببینید.



شکل ۵-۳: نمایش اولیه از ورودی U90223 در GenBank

۵. از منوی Send To گزینه Text را انتخاب کنید، تا بدینوسیله ورودی مورد نظر را به حالت text یا flat-file در آورید.
۶. سپس نتایج بدست آمده را با انتخاب گزینه Save As از منوی File ذخیره نمائید.

اگرچه این انتخاب مربوط به ژن انسانی است، ولی ورودی U90223 در GenBank با ورودی X01714 (که توصیف کننده همولوگ باکتریایی آن می باشد) تفاوت چندانی ندارد. در قسمت با لایی ورودی، کلید واژه های مربوط به اطلاعات کلی شامل LOCUS، ACCESSION، DEFINITION و VERSION به ترتیب ردیف شده اند.

خط KEYWORD که عبارات مربوط و قابل جستجو را لیست می کند، در مورد ورودی U90223 خالی می باشد. متأسفانه این اتفاق تصادفی نیست، بلکه یک مشکل رایج بانک های اطلاعاتی توالی ها می باشد. در نتیجه جستجو های بانک های اطلاعاتی بر اساس واژگان کلیدی، معمولا قادر به بازیابی تمامی توالی های مرتبط در GenBank نیست. همچنین، اطلاعات بخش های SOURCE و REFERENCE اغلب از دست رفته و ناقص می باشند. برای مثال در ورودی U90223، بیان شده

که توالی موجود از منابع منتشر نشده است. اما یک جستجو ساده در PubMed (به فصل ۲ مراجعه نمایید) با استفاده از اسامی نویسندگان (Ladner RD و Caradonna SJ) مقاله ی مورد نظر را که در سال ۱۹۹۷ !در مجله Biological chemistry چاپ شده است، را به ما نشان خواهد داد. توجه به این امر ضروری است که شما نباید انتظار به روز بودن GenBank و یا هر بانک اطلاعاتی دیگر را داشته باشید. از طرفی تعجب نکنید که در بعضی مواقع بازیابی بعضی توالی ها به وسیله جستجو در PubMed امکان پذیر نمی باشد.

در شکل ۶-۳ می توانید قسمت FEATURES از ورودی U90223 را مشاهده کنید. کلید واژه CDS، توالی منطقه کدکننده (۸۲۱-۶۳) مربوط به نوع میتوکندیایی dUTPase انسانی است. اگر فرایند فرضی ترجمه ORF به آمینواسید را از دنبال کنید، کلید واژه *sig peptide* مکان توالی ای که مسئول هدفمند کردن پروتئین برای میتوکندری است، را نشان می دهد، و کلید واژه *mat peptide* حدود دقیق پپتید نهایی را برای شما فراهم می سازد.

در این مورد خاص ورودی GenBank انسانی نسبت به نوع باکتریایی آن خیلی هم پیچیده نیست، زیرا ما به دقت ورودی را انتخاب کرده ایم که یک توالی mRNA را توصیف می کند، نه یک توالی ژنومی. در سطح mRNA بالغ، ارتباط بین پروتئین توالی های نوکلئوتیدی کد کننده آن، در یوکاریوت ها با این ارتباط در پروکاریوت ها معادل است. در قسمت بعدی به شما نشان خواهیم داد که اطلاعات مربوط به ژنوم خالص می تواند کمی پیچیده تر و گمراه کننده تر باشد.

درک مفهوم ورودی یک ژنوم یوکاریوتی در GenBank

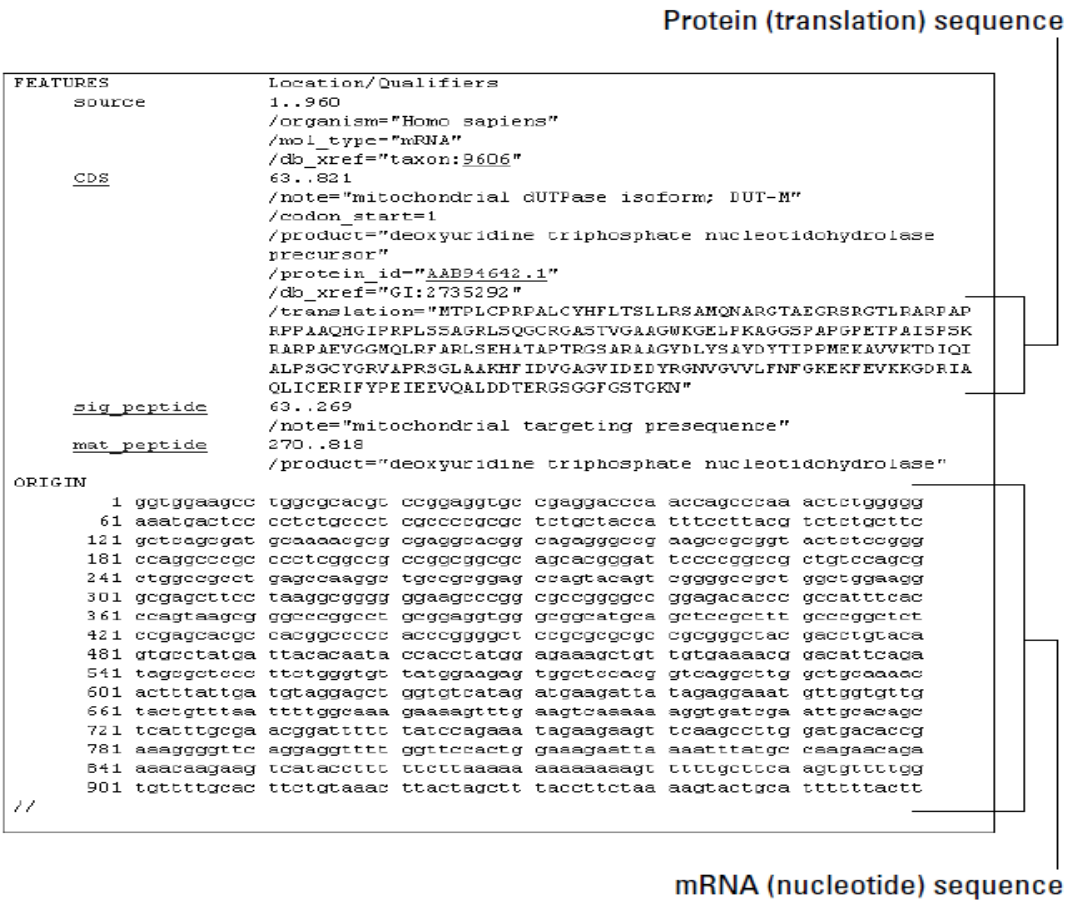
در قسمت قبل ورودی از GenBank که مربوط به توالی mRNA ژن dUTPase انسانی است، شرح داده شد. در این جا می خواهیم به ورودی AF018430 در GenBank که مرتبط با توالی ژنی که mRNA مربوطه از آن ساخته شده است، نگاهی بیاندازیم.

در ابتدا مجبور به استفاده از همان دستورالعمل قبلی هستیم :

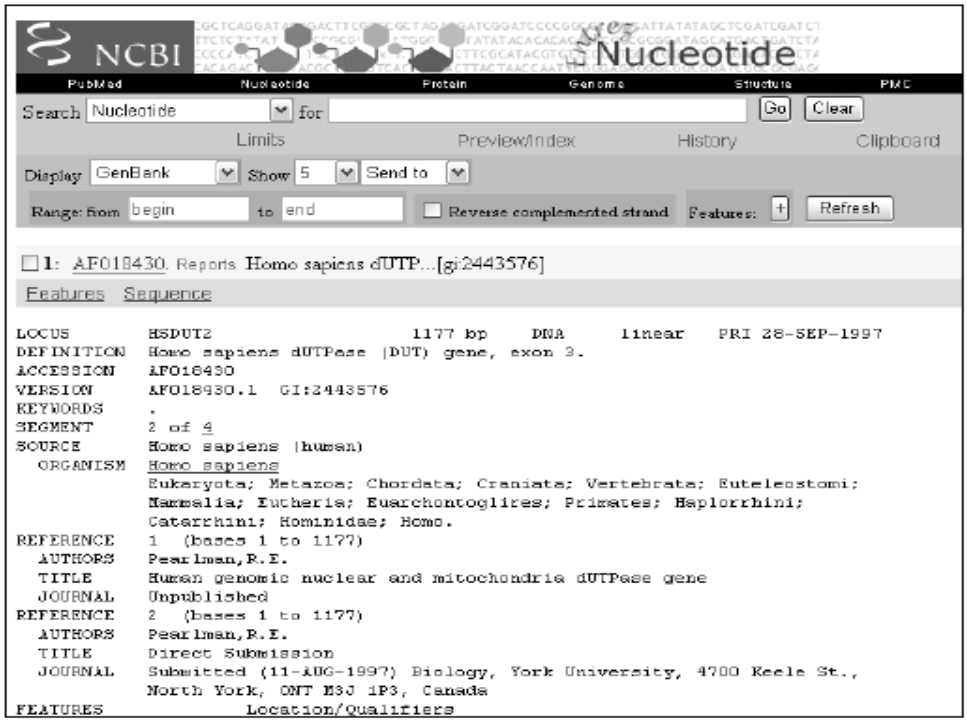
۱. به آدرس اینترنتی www.ncbi.nlm.nih.gov/entrez/ بروید. صفحه اصلی NCBI PubMed ظاهر می گردد.
۲. از منوی Search، گزینه Nucleotide را انتخاب کنید.
۳. در پنجره ی For، عبارت AF018430 را تایپ کرده و بر روی گزینه Go کلیک نمائید.
۴. بر روی لینک AF018430 کلیک نمائید.

اطلاعات ورودی AF018430 همانطور که در شکل ۷-۳ مشخص است، ظاهر می گردد.

اکنون برای اولین بار قادر به خواندن اطلاعات یک ورودی ژنوم انسانی در GenBank هستید. برای شروع یک ورودی ساده را انتخاب نمودیم، ولی با این حال این این ورودی شامل مواردی است که تنها در ورودی های یوکاریوتی با آن مواجه می شوید.



شکل ۶-۳: قسمت قسمت Features از ورودی U90223



شکل ۷-۳: ورودی AF018430 در GenBank

قسمت فوقانی این ورودی شامل اطلاعات عمومی است که با کلید واژه های زیر ارائه شده اند:

- ✓ **LOCUS:** نام Locus در اینجا HSDUT2 می باشد. ادامه این خط بیان کننده این مطلب است که با یک DNA خطی با ۱۱۷۷ جفت باز مواجه هستیم.
- ✓ **DEFINITION**: این قسمت به ما نشان می دهد که نام ژن مربوطه DUT بوده و ورودی مربوط به اگزون ۳ می باشد. به یاد داشته باشید که در ژن های انسانی، مناطق کدکننده پروتئین به صورت منفصل قرار گرفته اند، و به آنها اگزون گفته می شود.
- ✓ خطوط **ACCESSION**، **VERSION** و **KEYWORDS** استاندارد می باشند. توجه نمایید، در ورودی AF018430، خط KEYWORDS خالی است، یعنی شما نمی توانید با جستجوی محدود شده به فیلد KEYWORDS این ورودی با پیدا کنید.

- ✓ **SEGMENT:** این قسمت به ساختار موزائیکی ژن های یوکاریوتی اشاره دارد. در این قسمت به این نکته پی خواهید برد که ورودی GenBank حاضر، دومین قسمت از یک ورودی بزرگتر ۴ قسمتی است. لازم به ذکر است که شما به همه این ۴ قسمت برای بازسازی توالی کامل mRNA، که به عنوان الگویی برای ساخت پروتئین بکار می رود، نیاز دارید.
- ✓ خط های **ORGANISM**، **SOURCE** و **REFERENCE** استاندارد می باشند.

جدول FEATURES در حقیقت آن چیزی است که یک ورودی ژنوم یوکاریوتی را خاص و ویژه می سازد، به همین دلیل نسبت به نوع پروکاریوتی آن طولانی تر است. این جدول شامل موارد زیر است (شکل ۸-۳):

- ✓ قسمت **source** شامل یک بخش به نام */map*/می باشد. در مورد ورودی AF018430، این بخش مشخص می کند که توالی مرتبط با کروموزم ۱۵ است، و اگر بخواهیم دقیق تر بگوییم، بر روی بازوی بزرگ این کروموزوم (q)، و درون باند سیتوژنتیکی[1] 1.1 q2 قرار گرفته است.
- ✓ کلید واژه **gene**، فرمول های به نظر پیچیده ای را معرفی می نماید. هدف این فرمول ها، توصیف صریح و روشن بازسازی mRNA های گوناگونی است که در ورودی های مختلف قرار گرفته اند.

به دقت به اولین فرمول ژنی توجه کنید تا بفهمید چگونه کار می کند:

```
AF018429.1:<1..1735,1..1177,AF018431.1:1..45,
AF018432.1:658..732,AF018432.1:884..954,
AF018432.1:1391..>1447
```

این فرمول در حقیقت چیزی نیست جز یک دستوالعمل متصل کننده اگزون ها، که چگونگی بازسازی mRNA را بیان می کند:

۱- نوکلئوتید های ۱ تا ۱۷۳۵ را از ورودی AF018429 استخراج نمایید.

۲- نوکلئوتیدهای ۱ تا ۱۱۷۷ ورودی AF018430 را اضافه نمایید.

[1] sytogenetic band

۳- نوکلئوتید های ۱ تا ۴۵ ورودی AF018431 را اضافه نمایید.

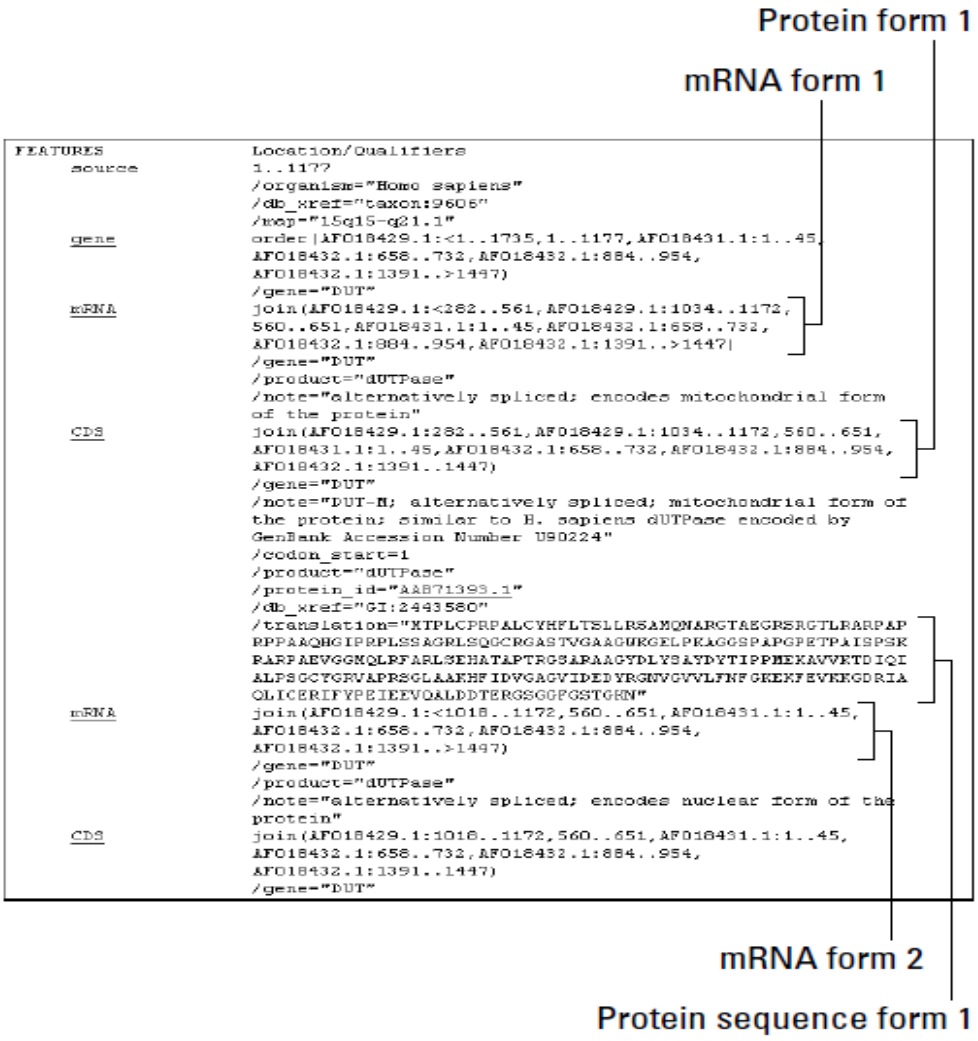
۴- نوکلئوتید های ۶۵۸ تا ۷۳۲ ورودی AF018432 را اضافه نمایید.

۵- نوکلئوتید های ۸۸۴ تا ۱۴۴۷ را از همان ورودی یعنی AF018432 اضافه نمائید.

۶- نوکلئوتید های ۱۳۹۱ تا ۱۴۴۷ را دوباره از همان ورودی اضافه کنید.

علامت کوچکتر (<) در ابتدای فرمول نشاندهنده این است که ممکن است ژن درست از قبل از موقعیت نشانگذاری شده شروع شود، در حالیکه علامت بزرگتر (>) در انتهای فرمول نشاندهنده این است که ژن مورد نظر احتمالا فراتر از موقعیت نشانگذاری شده ادامه می یابد.

✓ **The mRNA:** ورودی AF018430 از دوناحیه mRNA تشکیل شده است که جالب توجه است، زیرا این دو نشان می دهند که چگونه GenBank الگوهای بهم متصل کننده یک در میان[1] را نمایش می دهد. اتصال یک در میان، یکی از صفات ویژه ی بیان ژن در یوکاریوت ها است. اگر از منطق تجزیه مشابه آنچه برای ژن گفته شد، استفاده می کنید، می توانید نتایجی مشابه آنچه در جدول ۱-۳ خلاصه شده بدست آورید.



شکل ۸-۳: ورودیAF018430 در GenBank: جدول FEATURES برای یک ژن یوکاریوتی

[1] alternative splicing patterns

Mrna	*AF018429*	*AF018430*	*AF018431*	*AF018432*
Type 1 (Mitochondria)	282-561 1034-1172	560-651	1-45	658-732 884-954 1391-1447 ->
Type 2 (Nuclear)	<1018- 1172	560-651	1-45	658-732 884-954 1391-1447 ->

جدول ۱-۳: dUTPase mRNA متعلق به میتوکندری و هسته در ورودی AF018430

جدول ۱-۳ شما را در فهم بهتر آنچه در ادامه می خوانید کمک می نماید:

- دو نوع mRNA قابل جایگزینی وجود دارد : نوع ۱ (برای میتوکندری) ونوع ۲ (برای هسته)
- mRNA هسته، اگزون اول را (که در ورودی AF018430 ذخیره شده است)، ندارد، در حالیکه mRNA میتوکندری حاوی این اگزون است.
- mRNA هسته قسمتی از اگزون دوم (که در ورودی AF018429 ذخیره شده است) را، با یک قالب ترجمه متفاوت از همتای میتوکندریای آن، استفاده می کند.
- توالی های دو پروتئین در اگزون سوم با یکدیگر مشابه می شوند. این تشابه در دو فیلد */translation* (ETPAISPSK و غیره) دیده می شود. در شکل ۹-۳ می توانید توالی پروتئین هسته را مشاهده نمایید.

✓ کلید واژه Exon موقعیت اگزون منحصر به فرد موجود در این توالی را نشان می دهد. این اگزون در نزدیکی انتهای قسمت FEATURES در ورودی AF018430 قرار دارد (در شکل ۹-۳ به موقعیت های ۶۵۱-۵۶۰ توجه نمایید). اگر می خواهید وقوع چندین اگزون در یک ورودی را مشاهده کنید، ورودی AF018432 را ببینید.

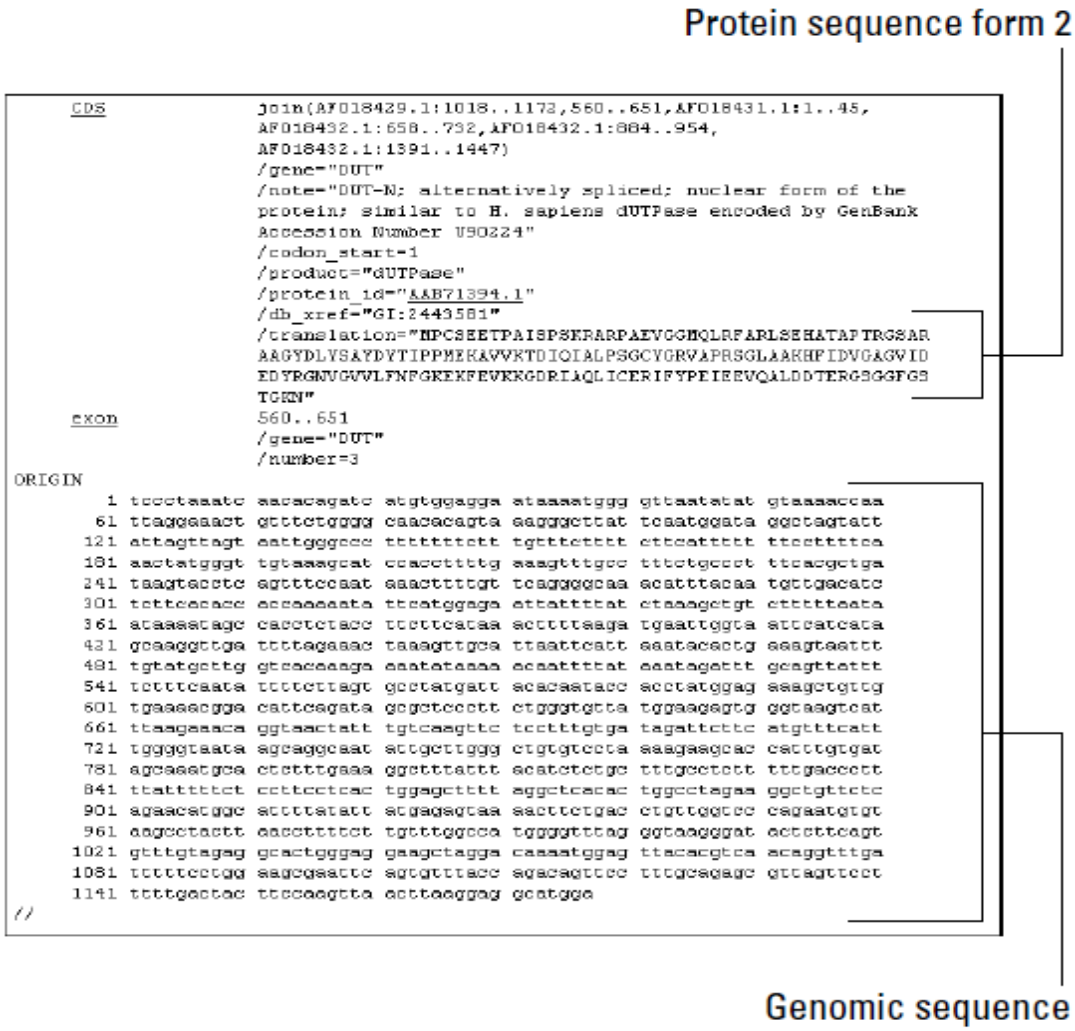
بخش توالی، در قسمت پایینی ورودی قرار دارد (شکل ۹-۳).

دشوارترین بخش این فصل، درک چگونگی اتصال قطعات توالی نوکلئوتیدی گوناگون، از ورودی های جداگانه GenBank، و تشکیل یک mRNA و مناطق کدکننده مرتبط به آن می باشد. اگر موفق به انجام این کار شده اید، به شما تبریک می گوییم. اکنون می توانید با خیال راحت کار خود را در GenBank ادامه دهید.

کار با ورودی های وابسته به هم در GenBank

در بسیاری از نمونه هایی که تا کنون استفاده کرده ایم، برای هر ورودی یک شماره دسترسی وجود داشت، برای مثال : U90223، X01714 و AF018430. در حالت عادی شماره های دسترسی را با مطالعه مقالاتی که به توالی مربوطه اشاره کرده اند و شماره دسترسی آنرا نیز صریحا ذکر کرده اند، می توان به دست آورد. این استراتژی برای شروع کار خوب است، ولی پس از دسترسی به اولین ورودی GenBank مرتبط با کار خود، می توانید سایر ژنهای مرتبط را از طریق انتخاب گزینه Related Sequences از منوی ظاهر شده پس از کلیک بر روی لینکی به نام Links که در قسمت راست صفحه مربوط به هر ورودی GenBank واقع شده است، پیدا کنید(شکل ۵-۳). بعنوان مثال اگر بر روی لینک مذکور در ورودی U90223 کلیک کنید، ۳۷ ورودی GenBank مرتبط با ژن انسانی dUTPase به دست خواهد آمد. این ورودی ها شامل

اشکال متفاوتی از mRNA و توالی های ناقص و نیز توالی هـای ژنـومی نـاقص (در اطـراف اگـزون هـا)، و دو تـوالی بـزرگ (154kb و192kb) از ناحیه ژنومی 15q21.1 می باشند. چندین توالی مشابه نیز وجود دارند.



شکل ۹-۳: ورودی AF018430 در GenBank : بخش پایینی یا همان بخش توالی

دسترسی به ورودی های بدون شماره های دسترسی در GenBank

اگرچه GenBank بهترین بانک اطلاعاتی برای جستجو بر اساس واژه کلیدی نمی باشد (اگـر دنبـال روش بهتـری مـی گردید، به بخش "استفاده از یک بانک اطلاعاتی ژن محور" در ادامه همین فصل مراجعه کنید)، ولی علاوه بر روش استفاده از شماره های دسترسی، جستجو در GenBank با استفاده از کلید واژه های gene و یا protein نیز ممکن است. تصور کنیـد می خواهید توالی نوکلئوتیدی کدکننده ژن انسانی dUTPase را پیدا کنید، اما شـماره دسترسـی آن را در اختیـار ندارید؛ در این صورت به روش زیر عمل کنید:

۱. به آدرس اینترنتی www.ncbi.nlm.nih.gov/entrez/ مراجعه نمایید. صفحه اصلی NCBI PubMed باز خواهد شد.

۲. از منوی Search، گزینه Nucleotide را انتخاب نمایید

۳. در پنجره جستجو human [organism] AND dUTPase [Protein name] را تایپ نمایید. این جستجو ۵ ورودی GenBank را در اختیارتان قرار خواهد داد: AH0055 8، AF018429، AF018430، AF018431، AF018432. این ورودی ها هر یک شامل انواعی از اگزون ها هستند: ورودی AH0055 8 اگزون ۱ و ۲، ورودی AF018430 اگزون ۳، ورودی AF018431 اگزون ۴ و ورودی AF018432 شامل اگزون ۵، ۶ و ۷ می باشد. ۴ ورودی آخر مشخص کننده توالی کامل آمینواسیدی پروتئین dUTPase در دو شکل (هسته ای و میتوکندریایی) می باشند.

۴. در ورودی **AF018432** بر روی لینک Links در سمت راست خط ID کلیک کرده و سپس گزینه Related Sequences را از روی منوی ظاهر شده انتخاب نمایید. با این کار ۲۰ ورودی ظاهر می شوند. از میان این ورودی ها، تعدادی حاوی توالی های mRNA می باشند، مانند ورودی U90223.

نکته: ما هنوز به همه ورودی های مرتبط با ژن انسانی dUTPase در GenBank دسترسی پیدا نکردیم. اگر بعضی از ورودی های جدید نگاهی بیاندازید، متوجه خواهید شد که dUTPase نام اختصاری برای dUTP pyrophosphatase است. اجازه بدهید در جستجوی بعدی خود از این واژه های جدید استفاده کنیم.

۵. در پنجره جستجو این عبارت را تایپ کنید:

human [organism] AND “dUTP pyrophosphatase” [Title]

توجه داشته باشد که علامت نقل قول ضروری است. سپس بر روی گزینه Go کلیک نمائید.

دقت داشته باشید که ما در این قسمت جستجوی خود را تنها به Title محدود نموده ایم و منظور ما جستجوی dUTPase به عنوان یک پروتئین(مانند انچه در مرحله ۳ انجام دادیم)، نبوده است.

این جستجو ۲۰ ورودی GenbBank را در اختیارتان قرار خواهد داد که دقیقا مشابه آنچه قبلا می شناختید، نیستند! در حقیقت این مسئله یکی از مشکلات معمول در بازیابی تمامی ورودی های مرتبط با یک موضوع خاص است و علت آن بکارگیری غیر همسان واژگان هم معنی (چون dUTPase، dUTP pyrophosphatase، deoxyuridine triphosphatase) و فیلد ها است. حدود نیمی از ورودی های فعلی متعلق به توالی های کوچک، منفرد و ناقص mRNA بوده که به اصطلاح به آنها *Expressed Sequence Tags* (ESTs) می گویند.

برای خارج کردن توالی های EST از لیست، می توانید از جستجوی محدود شده که در فصل ۲ به آن اشاره شد، استفاده کنید:

۱. با کلیک بر روی لینک Limits در زیر پنجره جستجو، منوی **Limits setting** را از باز کنید.

۲. گزینه **Exclude ESTs** را انتخاب نمایید.

۳. به قسمت بالای پوشه باز شده رفته و بر روی دکمه Go کلیک نمایید.

فقط ۱۱ ورودی GenBank از سد این محدود کننده اختصاصی عبور خواهند کرد.

سعی کنید فیلدهای دیگر را نیز در این روش دستی جستجوی محدود شده در Genbank استفاده نمایید.

استفاده از بانک های اطلاعاتی ژن محور

اخیرا NCBI در کنار بانک اطلاعاتیGenBank، بانک های اطلاعاتی دیگری را با توانایی بیشتر برای جستجو بر پایه ژن ارائه نموده است. در فرایند جستجو بر پایه ژن، بجای بررسی تمام قطعات توالی های مرتبط به ژن مربوطه، در واقع یک سؤال مرتبط با یک ژن خاص پرسیده می شود. مزیت اصلی بانک اطلاعاتی ژن محور این است که نتایج حاصل از آنها اختصاصی تر از لیست بلند بالای ورودی های GenBank است، لذا برای زیست شناسان قابل درک تر است. در حقیقت شما کل داستان را یکباره می فهمید!

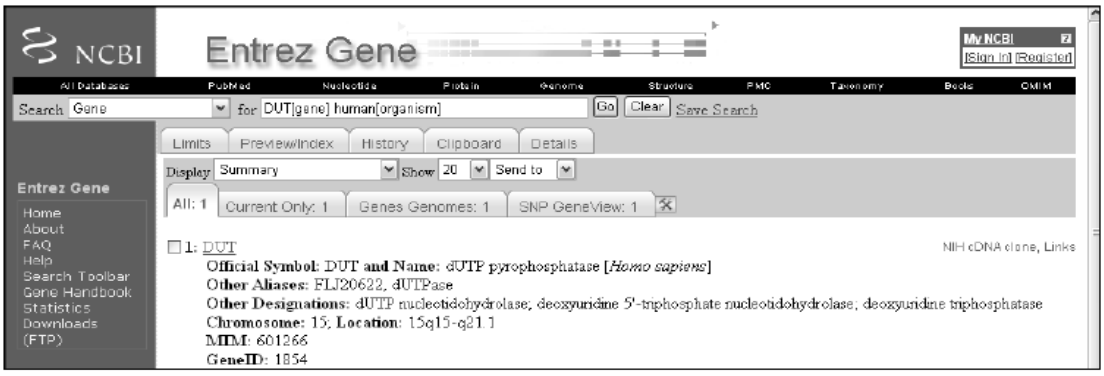
در این قسمت با Entrez/Gene، که در سرور NCBI موجود است، آشنا خواهید شد. این منبع امکان جمع آوری مطالب مهم و مرتبط با یک موقعیت ژنتیکی[1] (مکان اختصاصی یک ژن روی کروموزوم) را فراهم می سازد. با استفاده از این سرویس شما قادر خواهید بود به سرعت به تمامی اطلاعات موجود پیرامون ژن مورد نظر خود و یا اطلاعات ژنومی که ژن مورد نظر در آن قرار گرفته است، دست پیدا کنید.

۱- به آدرس اینترنتی www.ncbi.nlm.nih.gov/entrez/ مراجعه کنید. صفحه اصلی NCBI/Entrez ظاهر خواهد شد.

۲- از منیوی Search، گزینه Gene را انتخاب کنید.

۳- در پنجره For عبارت DUT [gene] human[organism] را تایپ کنید، سپس بر روی Go کلیک کنید (مطابق شکل ۱۰-۳) .

جستجو ی دیگری مثل DUT [gene name] AND "homosapiens" [organism] شما را به نتایج مشابهی خواهد رساند. مهمتر اینکه با انتخاب گزینه Gene از منوی Links در ورودی GenBank مربوط به نوکلئوتید با شماره دسترسی U90223، باز هم به همین نتیجه خواهید رسید(شکل ۵-۳).از بین روش های ذکر شده، بهترین راه ارتباط دادن یک قطعه از توالی با ژن مربوطه به آن است.



شکل ۱۰-۳: صفحه ورودی Entrez Gene برای ژن انسانی DUT

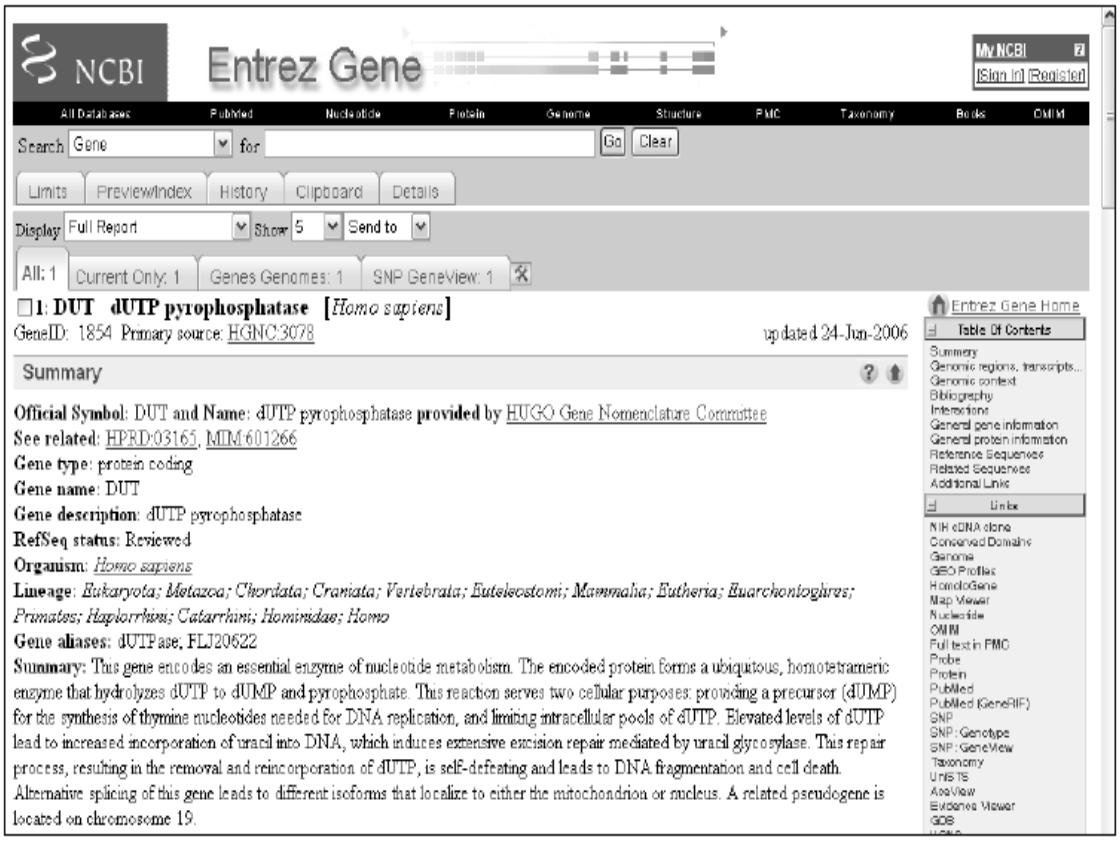
صفحه ی نتایج که در شکل ۱۰-۳ مشاهده می شود، راهی برای دسترسی به حجم زیادی از اطلاعات است. با کلیک بر روی لینک DUT (و یا با تغییر حالت Display به Full Report)، می توانید به اطلاعات گسترده ای پیرامون این ژن خاص و ژنومی که در آن قرار گرفته، دست یابید.

[1] genetic locus

در قسمت فوقانی ورودی DUT (شکل ۱۱-۳)، توضیح کلی پیرامون این ژن، عملکرد محصولات آن، و همچنین لینک های گوناگونی (در منوی سمت راست) که امکان دسترسی به بانک های اطلاعاتی و یا دیگر فایل های NCBI را برای شما فراهم می سازد، ارائه شده است.

شکل ۱۲-۳ قسمت بعدی این ورودی را نشان می دهد: یک شکل شماتیک از ساختار ژن DUT انسانی به همراه ۷ اگزون آن است که درون ناحیه ای به طول ۱۱۹۰۹ جفت باز در DNA ژنومی کروموزوم ۱۵ قرار گرفته اند.

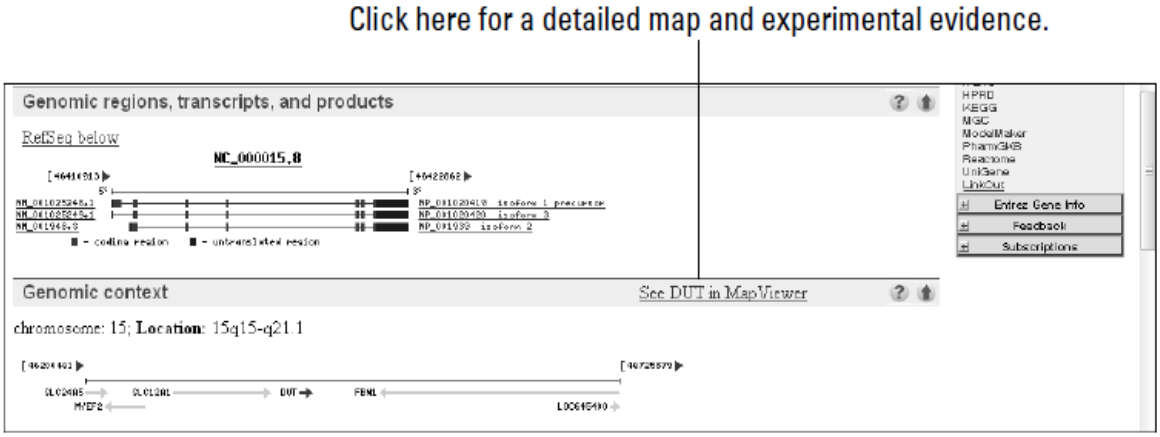
در ادامه این ورودی، بخش های اضافی حاوی اطلاعاتی پیرامون پتانسیل بر هم کنش با دیگر محصولات ژنها، توالی های همولوگ در دیگر ارگانیسم ها، عملکردهای پروتئین و راه های متابولیکی مرتبط با آن و همچنین لیستی از تمام ورودی های توالی مرتبط در GenBank را در اختیار شما قرار می دهد.



شکل ۱۱-۳: بخش بالایی ورودی ژن انسانی DUT

هر بخش، چندین لینک در اختیارتان قرار می دهد، که برای رسیدن به درک کاملی از ژن مورد نظر به شما کمک می کنند. اطلاعاتی که در GenBank می بینید، در واقع انواع توالی های mRNA مرتبط با ژن مورد نظر شما است که مشاهده و در این بانک اطلاعاتی ثبت شده اند. با مشاهده اطلاعات مربوط به توالی ها، حتما به این مطلب پی خواهید برد که تغییرات موجود به طور عمده به دو اگزون ابتدایی مربوط می شود. این انواع حاصل از تغییرات[1]، شامل انواع هسته ای و میتوکندریایی dUTPase می باشند(جدول ۱-۳).

[1] variants or alternative transcripts



شکل ۱۲-۳: متن ژنوم وجزئیات مربوط به mRNA های دیگر برای ژن انسانی DUT

به منظور مشاهده دقیق و جزئی ساختار ژن مورد نظر بر روی لینک See DUT in MapViewer کلیک نمائید. در صورت تمایل می توانید بحث های تجربی در مورد مدل های گوناگون mRNA را که در سمت راست و پایین نوکلئوتید واقع شده است، برای خودتان مرور کنید.

کار با بانک های اطلاعاتی ژنوم های کامل

جدید ترین بانک های اطلاعاتی ژن محور، در حقیقت در پاسخ به گسترش پروژه های توالی یابی کامل ژنوم از سوی بیوانفورماتیک مدرن ایجاد شده اند. هدف این منابع جدید جمع آوری تمامی اطلاعات مورد نیاز در مورد یک ارگانیسم مورد نظر، به صورت کاملا تفکیک شده، می باشد، بنابراین شما براحتی قادر خواهید بود که بر روی آنالیز تمامی ژنهای یک ژنوم خاص متمرکز شوید. این منابع جدید همچنین امکان مقایسه بین ژنوم های کامل را فراهم می کنند. این کار در واقع در قالب فیلد جدیدی با نام ژنومیک مقایسه ای[1] می باشد. اگر چه این قبیل بانک های اطلاعاتی قادر به فراهم سازی اطلاعات مربوط به توالی ژن های منفرد هم می باشند، قابلیت جدید و اضافی آنها بیان ارتباط بین ژنها در چارچوب یک ژنوم کامل است. بهترین راه ممکن برای کار کردن با این قبیل بانک های اطلاعاتی، این است که در این پایگاه ها در سطوح متفاوتی از جزئیات وارد شویم تا به سطح مطلوبی از جزئیات که پاسخگویی پرسشهای مورد علاقه ماست، برسیم. بانک های اطلاعاتی ژن محور همچنین آسانترین راه ممکن برای جمع آوری توالی های پروتئین یا ژن یک ارگانیسم را به سرعت و تنها با یک کلیک فراهم می کنند. کار خود را در بانک اطلاعاتی ژنوم کامل با نگاهی به بخش های ژنوم ویروس (در سایت NCBI) شروع می کنیم.

کار با ژنوم های کامل ویروسی

ویروس ها موجوداتی اسرار آمیز در گوشه ای از دنیای پیرامون ما هستند. این ماشین های بسیار هوشمند که از کوچک ترین قطعات مولکولی ممکن ساخته شده اند، تکثیر مولکول های نوکلئیک اسید (ژنوم ویروسی) را با استفاده از امکانات سلول های میزبان (یوکاریوت، باکتری و یا آرکئوباکتری ها)، تضمین می کنند. به طور معمول این موجودات ممکن است نادیده گرفته شوند و یا ممکن است موجب بیماری های بسیار وحشتناک و یا اپی دمی هایی چون آبله ، فلج اطفال و یا ایدز

[1] Comparative Genomics

شوند. ژنوم های ویروسی دراشکال مختلف بیوشمیایی چون : RNA/DNA ، خطی/کروی، تک رشته ای/دورشته ای و در سایز های مختلف از چند Kb تا یک میلیون جفت باز وجود دارند. مشاهده منابع مرتبط با ژنوم ویروسی در NCBI، یکی از بهترین راههای ممکن برای دسترسی به اطلاعات بیشتر در مورد آنهاست. برای نمونه، ما به شما نحوه دستیابی به اطلاعات عامل بیماری وحشتناک و مهلک ایدز (ویروس تیپ ۱ نقص سیستم ایمنی انسانی[1] و یا به اختصار HIV1) را نشان می دهیم:

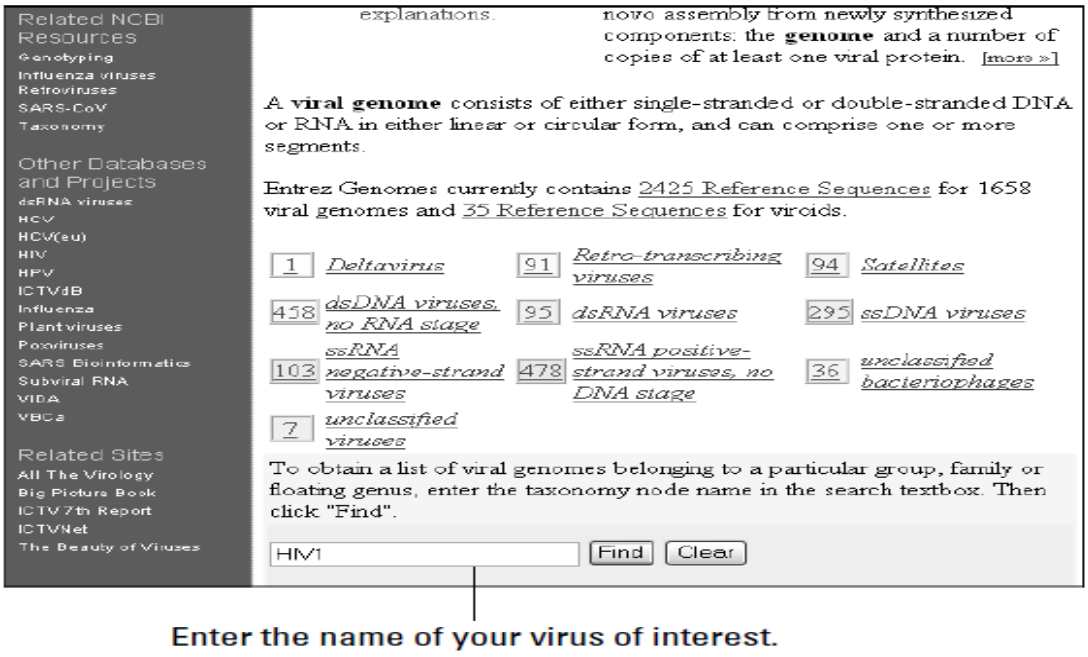
۱- به آدرس اینترنتی www.ncbi.nlm.nih.gov/entrez/ بروید. احتمالا دیگر آدرس این سایت را به خاط سپرده اید.

۲- به منیوی سیاه رنگ بالای پوشه رجوع کرده و روی گزینه Genome کلیک نمایید. با این کار صفحه Entrez Genome ظاهر می گردد.

۳- بر روی لینک Viruses در سمت راست پوشه کلیک نمایید. صفحه Viral Genomes reference ظاهر می گردد.

۴- صفحه مربوط به منابع ژنوم های ویروسی را تا پیدا کردن جدول توالی های ژنوم ویروسی پایین بیاورید. این جدول شامل دسته هایی چون دلتاویروس ها، رتروویروس ها و ... می باشد (شکل ۱۳-۳).

۵- در پنجره جستجو HIV1 را تایپ کنید و سپس بر روی دکمه Find کلیک نمایید. جستجوگر خلاصه ای کلی و زیبا از ژنوم HIV1 را، مشابه آنچه در شکل ۱۴-۳ می بینید، در اختیارتان قرار می دهد. در انتها شکلی با قابلیت کلیک کردن، مشخصات و موقعیت های نسبی تمامی ژنها را نشان می دهد.

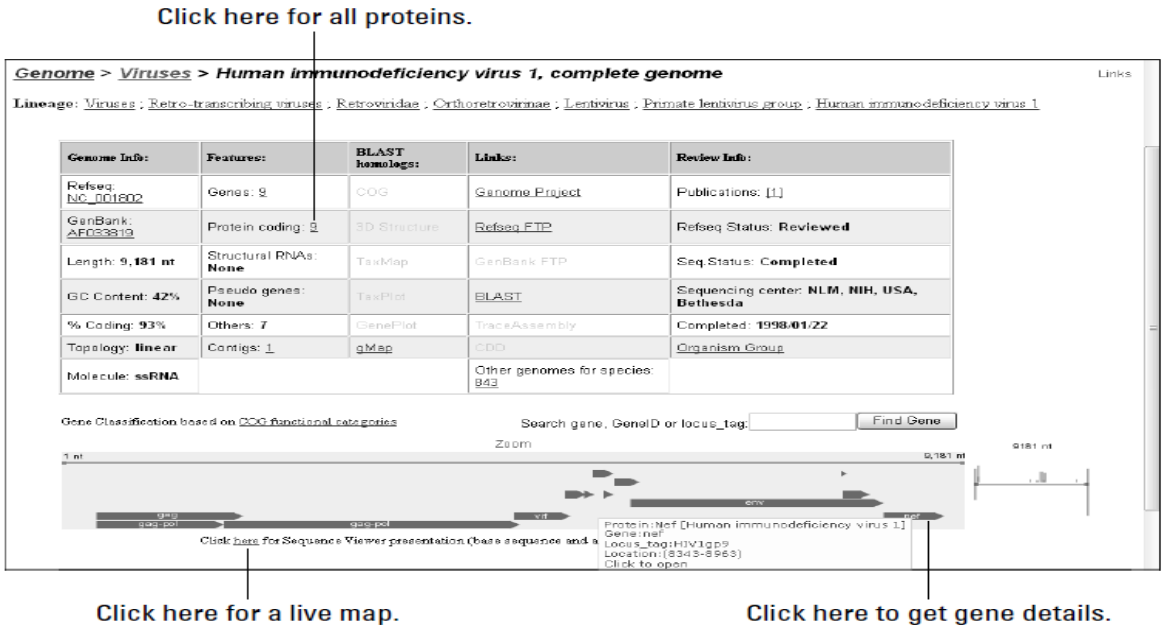


شکل ۱۳-۳: جدول مربوط به تمامی توالی های ژنوم ویروسی

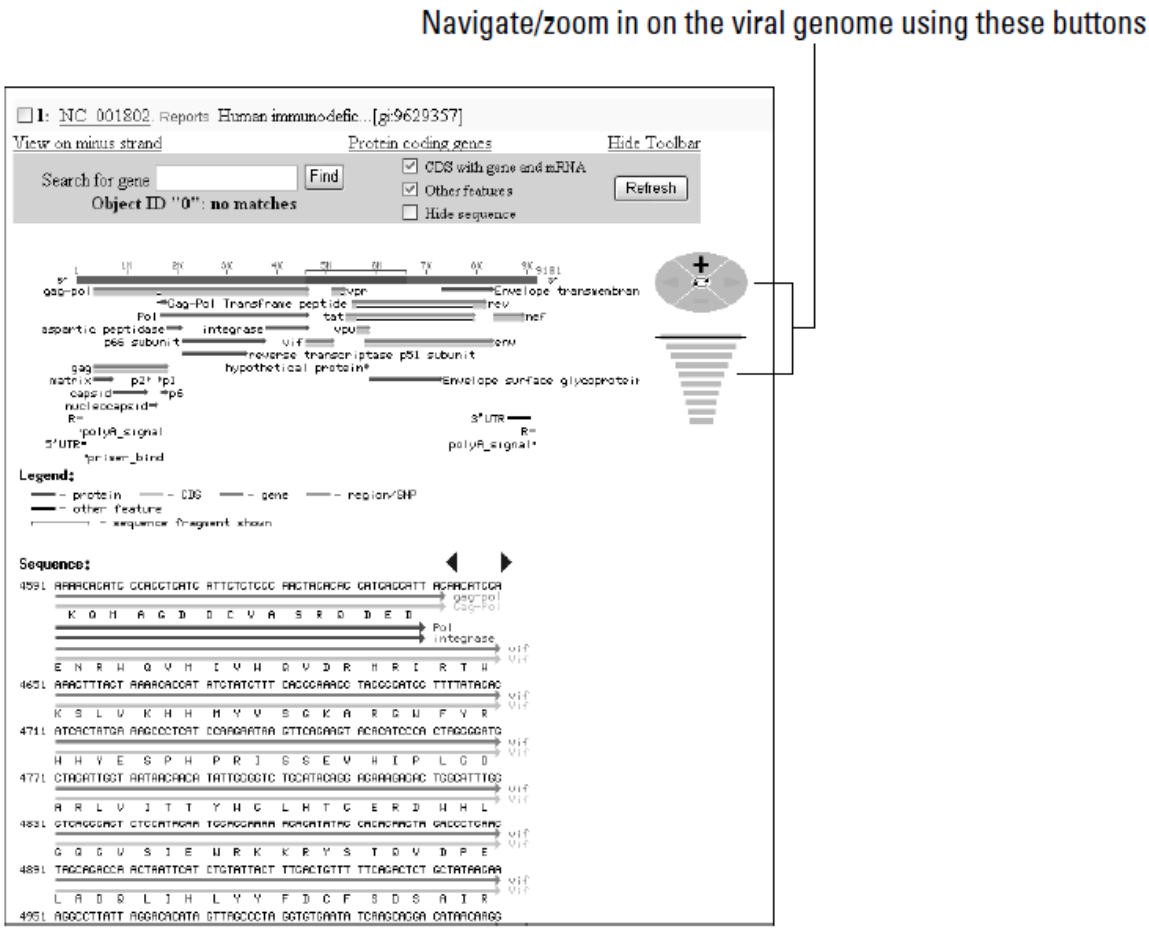
صفحات مشابهی برای تمامی ویروس ها صرف نظر از سایز آنها وجود دارد.

لینک here (در زیر شکل ۱۴-۳) نقشه ای زنده در زیر هر توالی نوکلئوتیدی، با قابلیت بزرگ نمایی و بررسی دقیق هر منطقه ژنوم، در اختیارتان قرار خواهد داد (شکل۱۵-۳).

[1] human immunodeficiency virus



شکل۱۴-۳: ساختار کلی ژنوم HIV-1



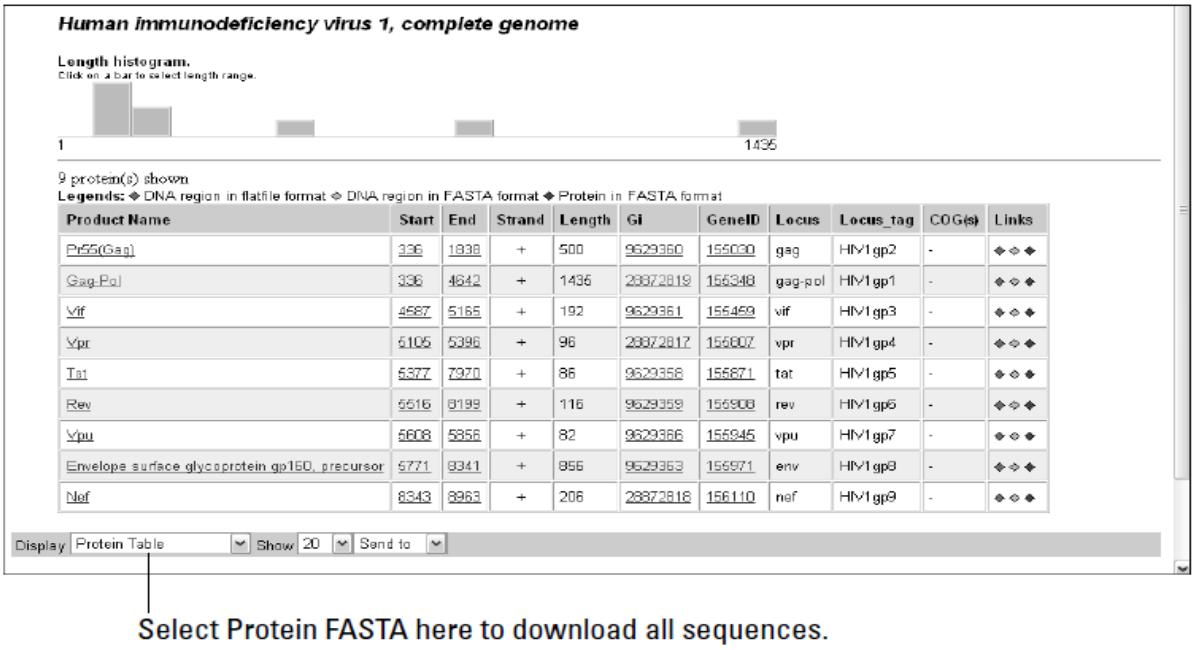
شکل ۱۵-۳: بزرگ نمایی بخشی از HIV-1 : همپوشانی Pol/Vif

شکل۱۵-۳ موقعیتی از ژنوم را نشان می دهد که در آن قسمت انتهایی ژن Pol با قسمت ابتدایی ژن Vif همپوشانی اندکی پیداکرده اند. اغلب ویروس ها ازیک توالی نوکلئوتیدی برای ساخت دو توالی آمینواسیدی متفاوت استفاده می کنند.

۶- بر روی گزینه Back در بالای صفحه کلیک نمایید. این اقدام شما را به ورودیHIV-1 Genome باز می گرداند(شکل ۱۴-۳).

۷- برروی شماره ۹ که بعد از عبارت Protein coding (ستون دوم، سطر دوم جدول) آمده، کلیک نمایید. صفحه لیست پروتئینی ظاهر می شود(شکل ۱۶-۳). با کلیک بر روی یکی از سه لوزی در ستون سمت راست، قادر به بازیابی توالی های پروتئین وDNA در فرمت های مختلفی چون GenBank و یا FASTA می باشید.

در صورت تمایل می توانید مجموعه کاملی از توالی های پروتئینی را با انتخاب Protein FASTA از منوی Display دانلود کنید.



شکل ۱۶-۳: دانلود توالی های ژن و پروتئین HIV-1

کار با ژنوم های کامل باکتریایی

NCBI Entrez یک ارتباط متقابل با تمامی توالی های ژنوم باکتریایی را دراختیارتان قرار می دهد. کافی است مطابق دستورالعمل زیر عمل نمایید:

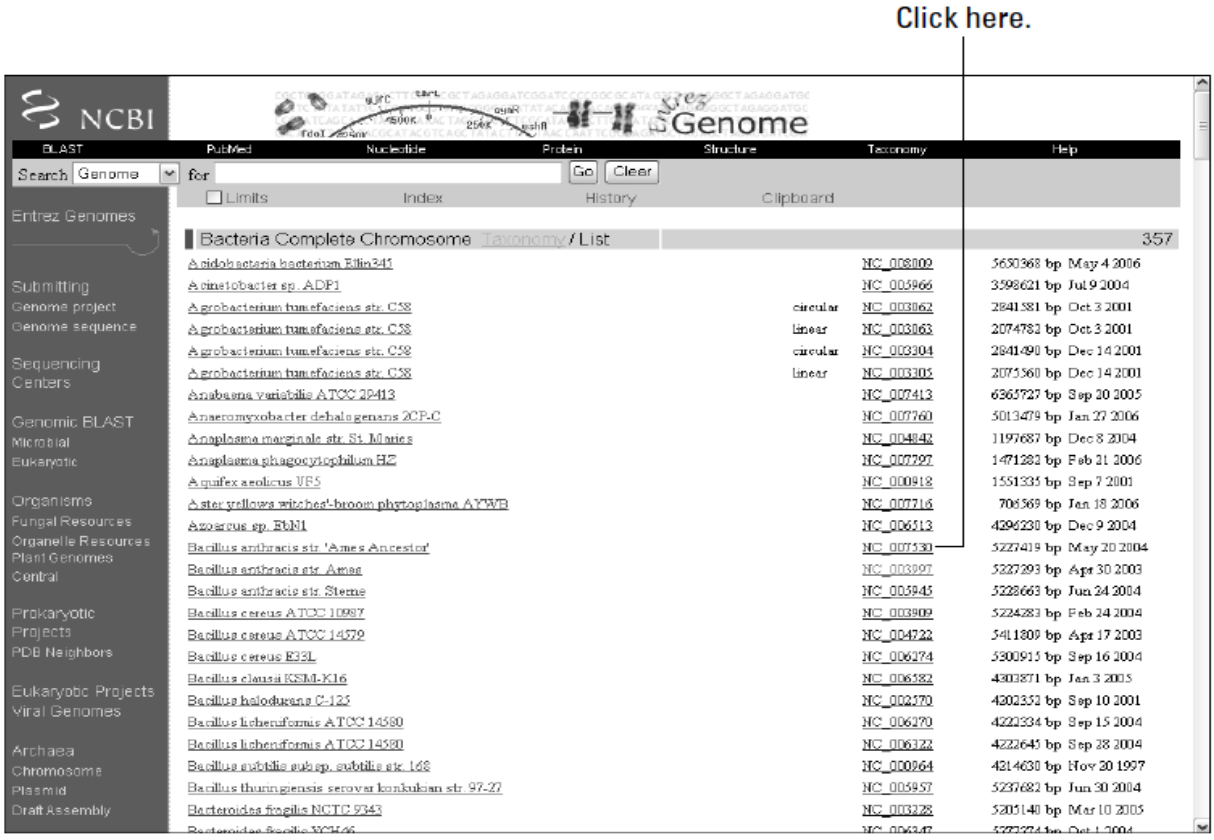
۱- به آدرس اینترنتی : www.ncbi.nlm.nih.gov/entrez/ بروید. صفحه اصلی NCBI PubMed ظاهر می گردد.

۲- بر روی گزینه Genome از منوی سیاه رنگی که نزدیک قسمت فوقانی صفحه قرار دارد، کلیک نمایید. صفحه Entrez Genome ظاهر می گردد.

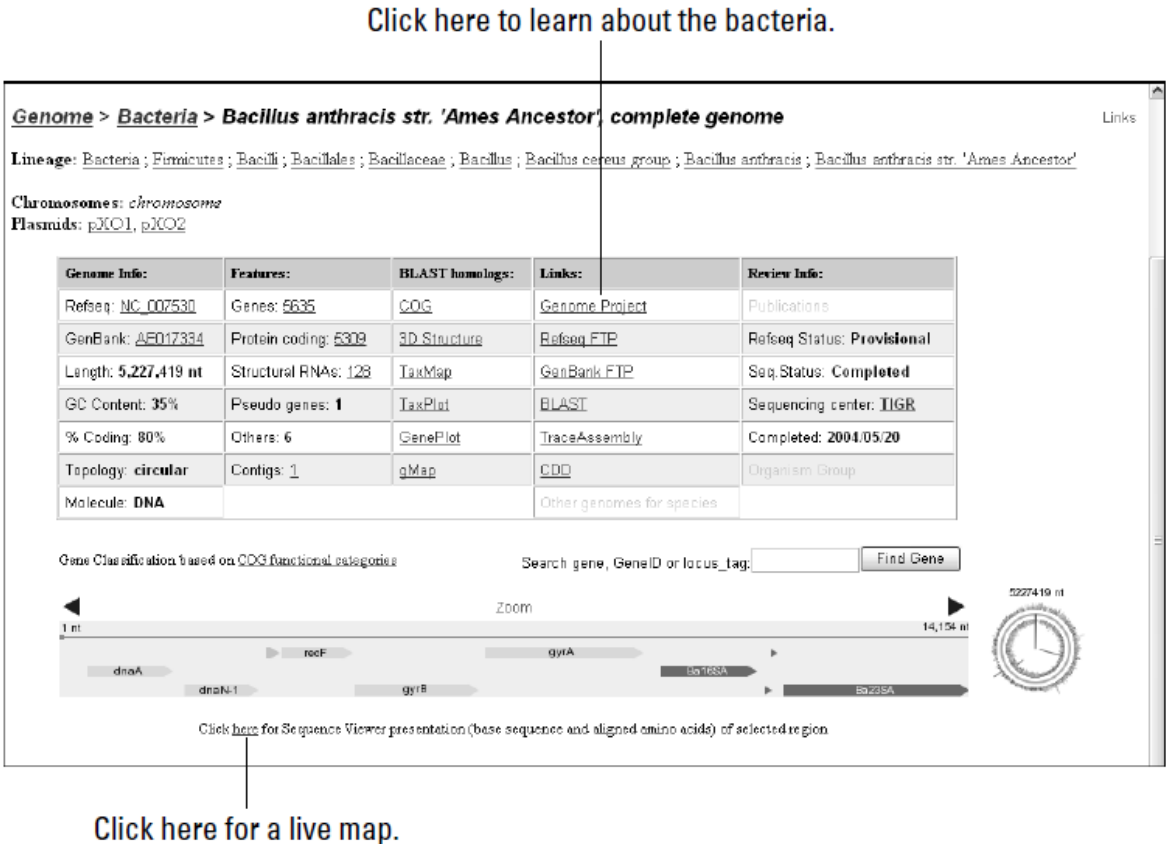
۳- در حاشیه آبی تیره در سمت چپ، بر روی لینک **Chromosome** که در زیر سرعنوان Bacteria قرار دارد، کلیک نمایید. لیستی از تمامی باکتری هایی را که کروموزوم های آنها بطور کامل توالی یابی شده اند ظاهر می شود. اگر مستقیما روی Bacteria کلیک کنید، یک لیست بلند تر که شامل قطعات DNA فرعی به نام پلازمید نیز هست، در اختیارتان قرار می گیرد.شکل ۱۷-۳ ابتدای این لیست را نشان می دهد.

۴- بر روی لینک NC_007530، که مربوط است به یک عامل بیوتروریسم خطرناک به نام ***Bacillus anthracis***. کلیک کنید. جستجوگر جدولی را در اختیارتان قرار خواهد داد که در آن محتویات و شاخصه های مهم ژنوم باکتریایی بصورت خلاصه بیان شده است. این در حقیقت مشابه آن چیزی است که در قسمت قبل در مورد ویروس ها بدست آوردید(شکل ۱۸-۳).

باکلیک بر روی لینک Here، به نقشه ای زنده مشابه آنچه در مورد ویروس HIV1 مشاهده کردید، دست خواهید یافت (به بخش قبلی مراجعه نمایید)، بااین تفاوت که ژنوم در اینجا طولانی تر است. در این نقشه قادر بزرگنمایی یک منطقه مورد نظر از ژنوم خواهید بود. جدول خلاصه ژنوم، حاوی لینک های فراوانی به آنالیزهای از پیش پردازش شده (عملکرد، تشابهات، روابط تکاملی) در مورد تمام ژنها می باشد،که می توانید آن را امتحان کنید.

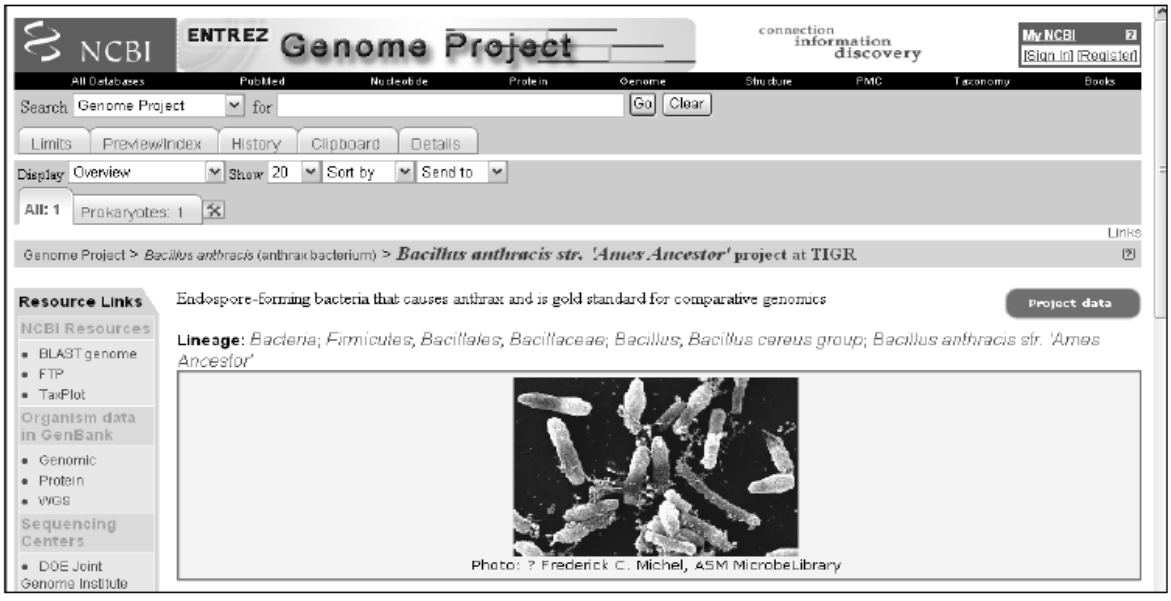


شکل۱۷-۳: لیستی از تمامی توالی های ژنوم باکتریایی در دسترس



شکل ۱۸-۳: خلاصه ژنومی *Bacillus anthracis* جد گونه Ames

۵- به جدول بازگردید(شکل ۱۸-۳) و بر روی لینک Genome Project کلیک نمایید. انجام این کار شرحی مختصر از باکتری، یک عکس، جزئیاتی پیرامون محل زیست آن و بیماری هایی که ایجاد می کند و نظایر آن در اختیار شما قرار می دهد. در این قسمت همچنین اهمیت کشف توالی های ژنوم آن را نیز بیان شده است.



شکل۱۹-۳: صفحه مربوط به پروژه ژنوم *Bacillus anthracis* جد گونه Ames

ژنوم های باکتریایی دیگر در TIGR

موسسه تحقیقات ژنوم[1] (به اختصار TIGR) مکانی است که در آن مجموعه ای از دانشمندان پیشگام در زمینه ژنومیک باکتریایی مشغول به کارند. در سال ۱۹۹۵ ، این دانشمندان برای اولین بار ژنوم کامل دو باکتری به نامهای *Hemophilus influenzae* و *Mycoplasma genitalium* راکشف نمودند. از آن پس، آنها به همکاری هم توانستند به بیش از ۷۰ ژنوم کامل باکتریایی دست یابند. TIGR یک سایت دارد که تقریبا مکمل NCBI است، زیرا این سایت علاوه بر ژنوم های کامل شده، اطلاعات مربوط به پروژه های توالی یابی ژنوم در حال انجام را نیز شامل می شود، از طرفی ابزارهای آنالیزگر خود را برای استفاده در این زمینه پیشنهاد می کند و نیز امکان جستجوی تشابهات موجود در توالی های ژنوم را، قبل از کامل شدن پروژه های توالی یابی آنها، در اختیار قرار می دهد. علاوه بر آن دانشمندان TIGR با استفاده از نرم افزار های خاص، دورنمائی از ژنوم هایی را که توالی یابی نموده اند، در اختیار کاربر قرار می دهند.

برای نگاهی اجمالی به آنها مراحل زیر را دنبال کنید:

۱. به آدرس اینترنتی www.tigr.org/tdb/ بروید. صفحه اصلی TIGR مطابق شکل ۲۰-۳ ظاهر می گردد.
۲. بر روی لینک Comprehensive Microbial Resource (CMR) کلیک نمایید. لیستی طویل از میکرواورگانیسم ها در مقابل شما ظاهر می گردد که از بین آنها می توانید میکروارگانیسم مورد نظر خود را انتخاب نمایید.
۳. یک میکروب را در این لیست انتخاب کرده و سپس در قسمت فوقانی صفحه مربوط به هر میکروب، ابزار های مختلف آنالیزگر را با کلیک بر روی لینک های Genome Searches،Genome Toolbox و Genome Analyses امتحان نمایید.

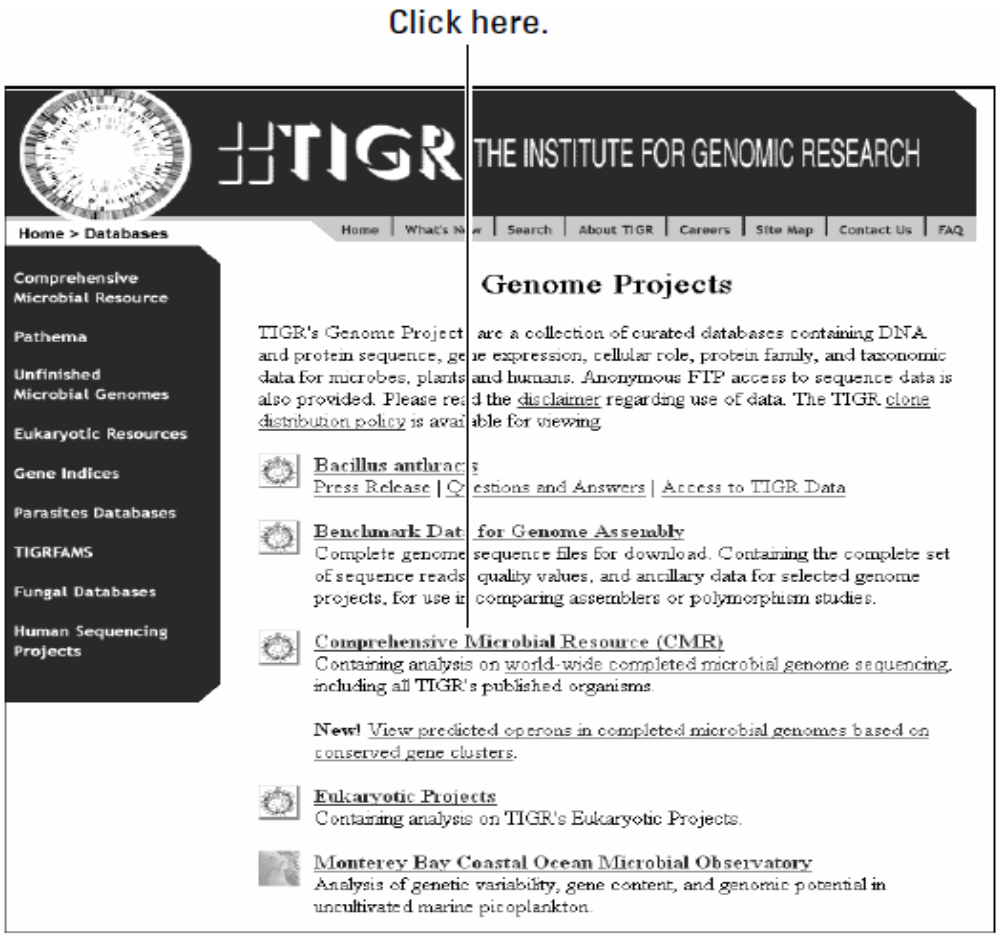
میکروب های موجود در محیط زیست در DoE

سازمان انرژی ایالات متحده[2] (DoE) نقش مهمی در زمینه ژنومیک میکروبی دارد. یک بخش الحاقی به آن به نام Joint Genome Institute به صورت تخصصی در زمینه ارگانیسم هایی که برای حفاظت و پاک سازی محیط زیست مهمند، یا دورنماهای جدیدی در زمینه رفع بحران انرژی (از جمله راههای ساده و راحت برای تولید هیدروژن) ارائه می کنند، فعالیت می کند.

به منظور دسترسی به اطلاعات این سازمان، مراحل زیر را دنبال کنید:

۱- به آدرس اینترنتی img.jgi.doe.gov/ مراجعه نمایید. صفحه اصلی منابع IMG[3] ظاهر می گردد (شکل ۲۱-۳). در اینجا نیز همانند NCBI و TIGR، مجموعه ای اختصاصی از ابزار های مورد نیاز در اختیارتان قرار می گیرد. مرحله بعدی، نگاهی است به صفحه نمایش ژنوم در IMG، که یکی از سودمندترین منابع IMG می باشد.
۲- به منظور دسترسی به صفحه نمایش ژنوم در IMG، ابتدا بر روی لینک Find Genomes کلیک نمایید. جدول ارگانیسم ها ظاهر می گردد.

[1] The Institute for Genome Reseach
[2] The U.S. Department of Energy
[3] Integrated Microbial Genemes



شکل ۲۰-۳: سایت Genome Projects در موسسه تحقیقات ژنوم (TIGR)

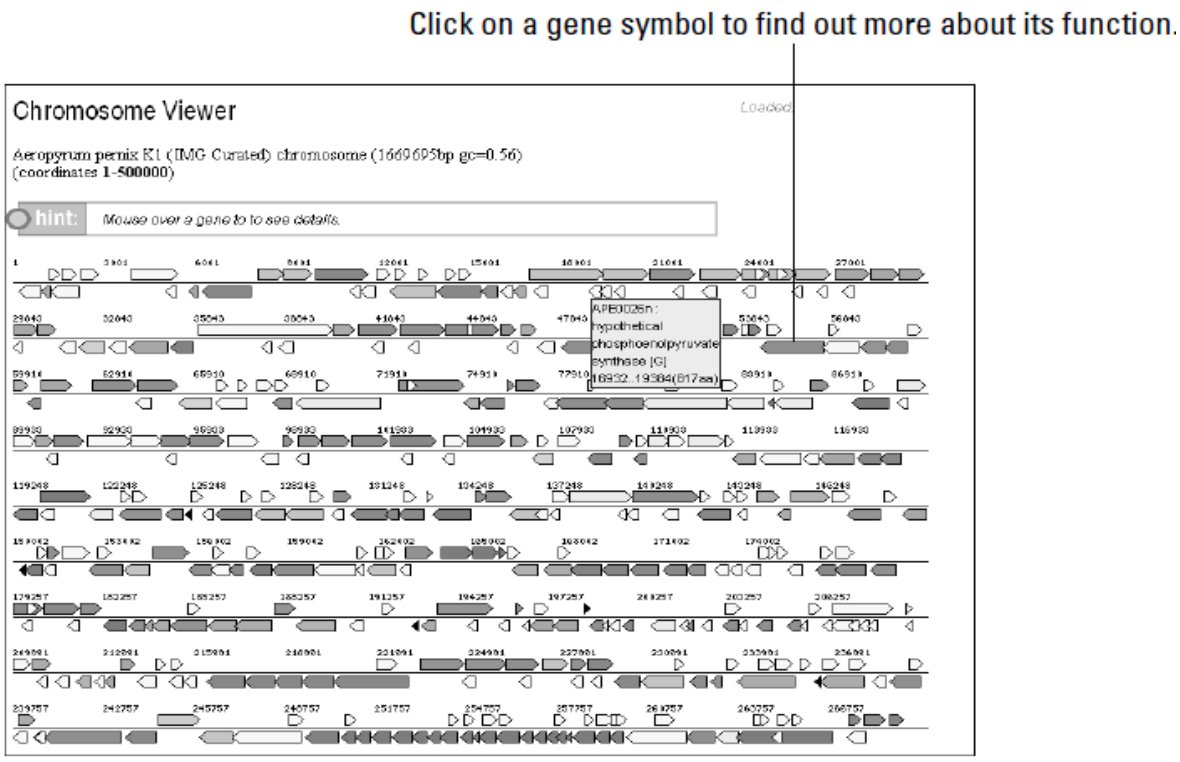
شکل ۲۱-۳: سایت IMG در DoE Joint Genome Institute

۳- ارگانیسم مورد نظر را با تیک زدن در مربع کنار آن و سپس کلیک بر روی گزینه Save Selections انتخاب کنید. به عنوان اولین ارگانیسم Aeropyrum pernix می تواند انتخاب خوبی باشد.

۴- در صفحه جدیدی که ظاهر می گردد،بر روی گزینه Compare Genomes کلیک کنید. صفحه ظاهر می شود.

۵- در صفحه Genome Statistics، در ردیف داربست DNA بر روی شماره ۱ کلیک نمایید.

۶- در صفحه جدیدی که ظاهر می گردد، رنج متناسب با کار خود را انتخاب نمایید (برای مثال 500000 . . . 1). با این کار یک نمایش زنده از محتویات ژنی میکروب مورد نظر در رنج مورد نظر حاصل می شود (شکل ۲۲-۳). با قرار دادن نشانگر موس در بالای علامت ژن، نام آن ظاهر می گردد و کلیک بر روی آن علامت، اطلاعات ارزشمندی را در رابطه با پروتئین مربوط به این ژن در اختیارتان قرار می دهد.



شکل ۲۲-۳: نمایشگر کروموزوم در منابع جامع ژنوم های میکروبی DOE

کشف ژنوم انسان

بدون شک توالی یابی ژنوم انسانی یکی از بزرگترین دستاوردهای علمی در عصر مدرن می باشد. از طرف دیگر دستکاری این ژنوم با بیش از ۳۰۰۰ میلیون نوکلئوتید موجود در ۲۳ کروموزم، کاری بسیار پیچیده است. مهمترین هدف پیش روی رشته بیوانفورماتیک، جمع آوری اطلاعات ژنهای انسانی از گذشته، حال و آینده در یک منبع قابل نگهداری و غیر پیچیده می باشد. برای درک صحیح اطلاعات ژنوم انسانی، بد نیست از وضع فعلی این اطلاعات با خبر باشید:

- ✓ اطلاعات مربوط به توالی نوکلئوتیدی ژنوم انسانی هم اکنون بطور کامل در دسترس است، البته به جز چند هزار جای خالی که در حال پرشدن است. دقیقا به همین علت است که مرتبا نسخه ها[1]ی جدید ژنوم انسانی منتشر می شوند. این جریان در مورد تمامی موجودات با ژنوم پیچیده نیز وجود دارد.

[1] versions or assemblies

- ✓ توالی ژنوم انسان بصورت خام بدست آمده است مرحله بعد تفسیر این اطلاعات و تهیه جدول FEATURES حاوی مشخصات جزئی و صحیح ژنوم انسان است.
- ✓ هر روز در سرتاسر جهان، اطلاعات جدیدی در مورد خصوصیات ژنوم انسانی و عملکلرد و نقش آن، با استفاده از گستره وسیعی از تکنیک ها، به دست می آید. اما یک نفر باید این اطلاعات را به طور کامل جمع آوری و بخوبی دسته بندی کند، و به صورت رایگان! در اختیار جامعه تحقیق و پژوهش قرار دهد.

فکر می کنید مورد آخر غیر ممکن است؟ عجله نکنید! چون ادامه مطلب در همین مورد است.

کسب اطلاعات بیشتر پیرامون پروژه Ensembl

صفحه اصلی Ensembl در اینترنت (www.ensembl.org) ، این اطلاعات را در اختیار شما قرار می دهد: Ensembl یک پروژه مشترک بین موسسه بیوانفورماتیک اروپا[1] (EBI) و مؤسسه Sanger می باشد که هر دو مؤسسه در نزدیکی Cambridge در انگلستان واقع شده اند. این دو مؤسسه سیستمی در قالب بانک های اطلاعاتی و نرم افزار برای پردازش و نگهداری اطلاعات به دست آمده از ژنوم های حیوانی به وجود آورده اند. البته لازم به ذکر است که این سیستم توجه خاصی به خویشاوندان نزدیک ما، یعنی مهره داران، دارد. این پروژه (مشابه دیگر پروژه هایی که در این بخش به آنها اشاره می کنیم) بطور جدی به همکاری طیف گسترده ای از آزمایشگاه ها و مراکز تحقیقاتی وابسته است. این پروژه به عنوان جزئی از پروژه های ژنوم انسانی شروع بکار کرده و با پروژه ژنوم موش کار خود را ادامه داده و اکنون شروع به بررسی سایر حیوانات نموده است. این نرم افزار ها و اطلاعات خام موجود بطور گسترده بین مراکز بیوانفورماتیک سراسر جهان در گردشند، که این مسئله خود می تواند زمینه ساز ایجاد یک پیچیدگی در ارتباطات متقابل شود.

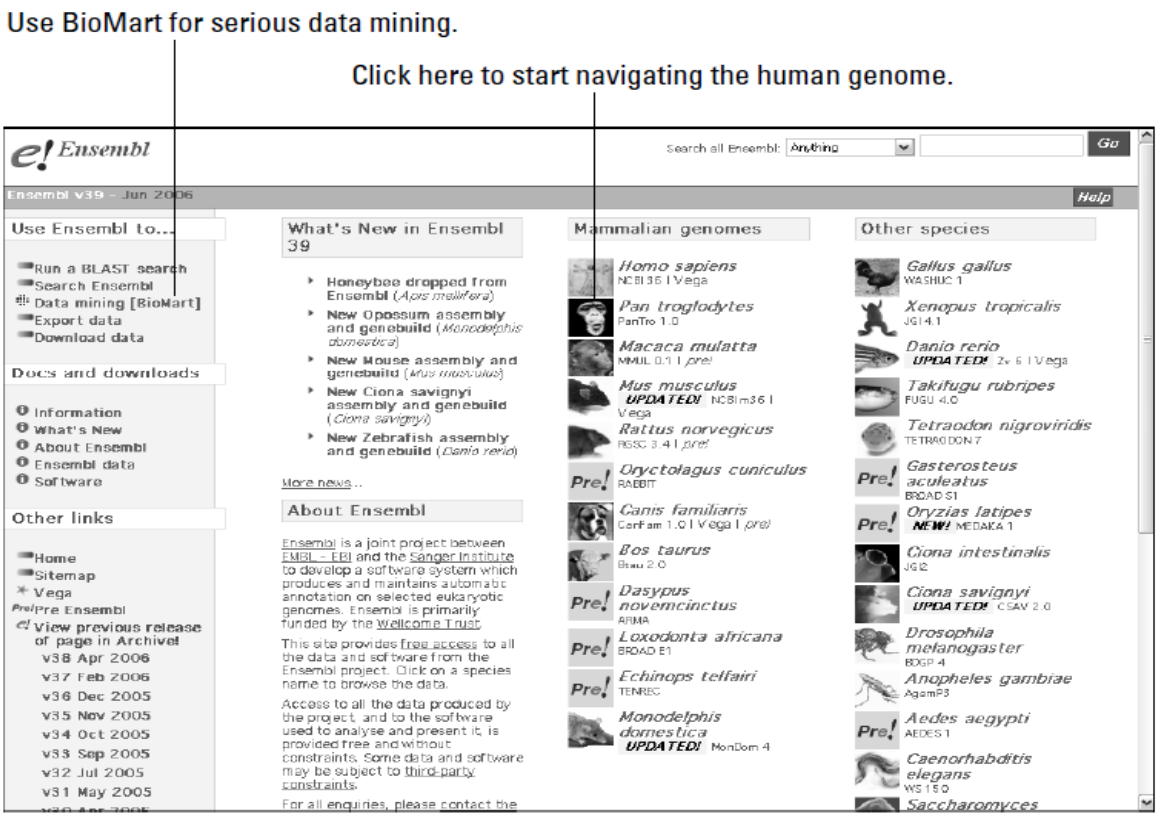
شروع کار با سایت Ensembl

با در نظر گرفتن پیچیدگی ژنوم انسانی، احتمالا دیگر نسبت به جستجو در منابع Ensembl و موضوعات بسیار متنوعی که با آنها روبرو می شوید، شگفت زده نخواهید شد. برای اینکه متوجه شوید چه در Ensemble چه گزینه هایی برای انتخاب دارید، بهترین راه این است که به سراغ قسمت راهنما در صفحه اصلی Ensembl بروید:

۱- به آدرس اینترنتی www.ensembl.org مراجعه نمایید(شکل ۲۳-۳).

صفحه ظاهر شده مملو از لینک های متنوع و فراوانی است، طوریکه توضیح بیشتر راجع به آن از بحث این بخش از کتاب خارج است. در صورت علاقه می توانید با صرف هفته ها وقت تمامی لینک های موجود در این سایت را امتحان نمایید، ولی اکنون بهتر است به نیمه راست این صفحه نگاهی بیاندازیم. در این قسمت لینک های ژنوم حیوانی بوسیله تصاویر کوچکی مشخص شده اند. در این قسمت نام گونه ها با نام لاتین معرفی شده اند؛ برای مثال شامپانزه (Pan troglodytes) ، خرگوش (Oryctolagus cunicumus) و یا آرمادیلو (Dasypus novemicinctus) . هدف اصلی ما در این قسمت بررسی ژنوم انسانی است، لذا بر روی گزینه DaVinci-inspired *Homo sapiens* کلیک نمایید.

1 European Bioinformatics Institute



شکل ۲۳-۳: صفحه اصلی پروژه Ensembl

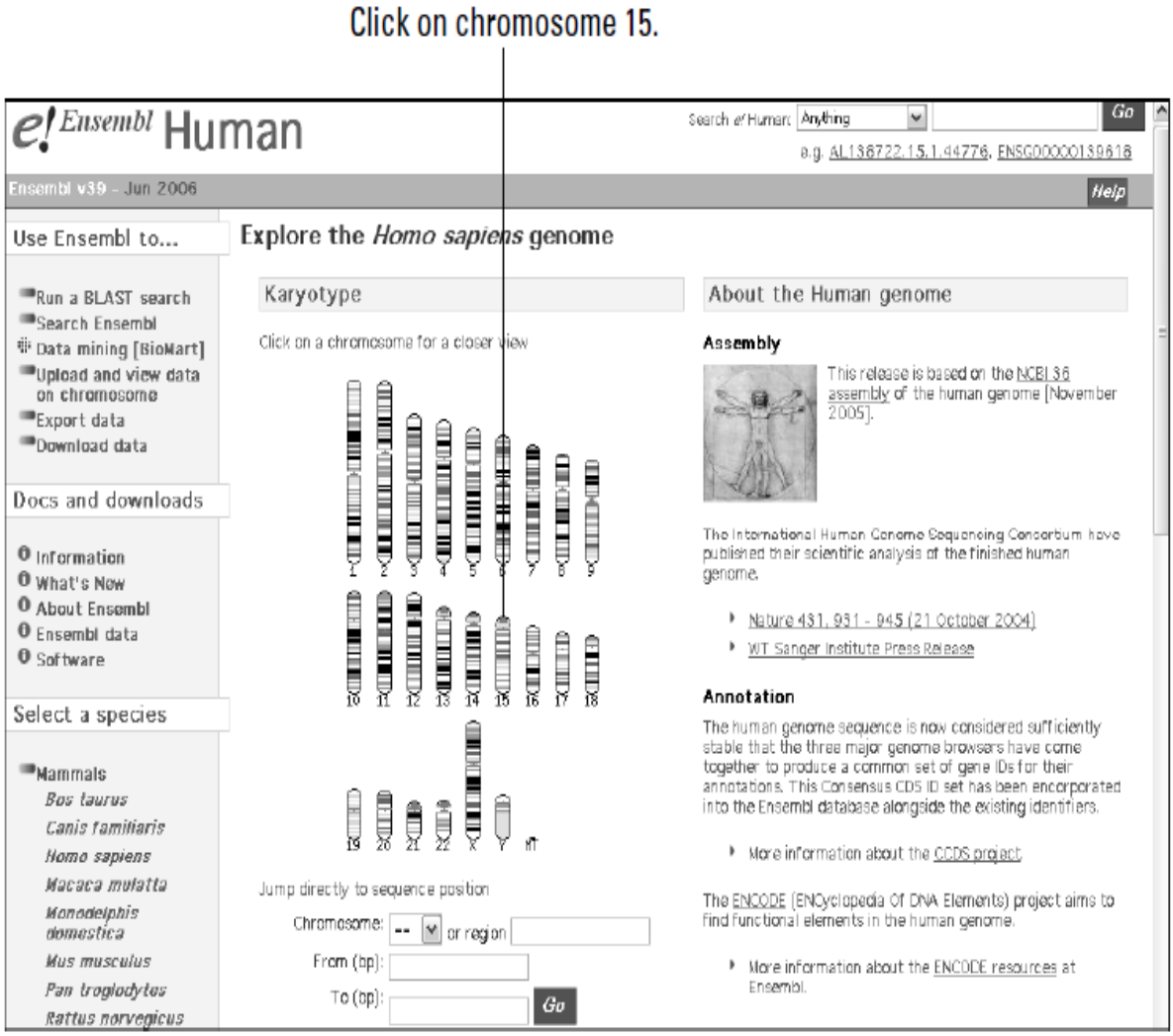
۲- روی گزینه Homo sapiens کلیک نمایید. صفحه ظاهر شده(شکل ۲۴-۳) یک شکل شماتیک از کروموزوم های انسانی مختلف را نشان می دهد(شماره گذاری بر اساس اندازه کروموزوم ها است)، و شامل کروموزوم های جنسی X و Y و به علاوه مولکول DNA میتوکندریایی انسانی نیز هست(میتوکندری : یک اندامک کوچک تولید کننده انرژی در سلول های انسانی). این شکل شماتیک به اصطلاح hot می باشد، به این معنی که در صورت کلیک بر روی یک ناحیه خاص صفحه جدیدی مرتبط با آن ناحیه در اختیار شما قرار می گیرد.

۳- بر روی شکل کروموزوم ۱۵ کلیک نمایید. با این کار زیرمجموعه ای از اطلاعات مرتبط با کروموزوم ۱۵ در اختیارتان گذاشته می شود(شکل ۲۵-۳). حال می توانید سؤالات تخصصی خود را بپرسید، مثلا: ژن dUTPase کجاست؟ راه سخت تر اینست که بر روی باند q21.1 کلیک کنیم و سپس ناحیه مورد نظر را به صورت دستی اسکن نماییم (برای یادآوری بیشتر به مطالب قبلی در این فصل و شکل ۸-۳ رجوع کنید). راه آسانتر استفاده از ردیاب Feature Locator در گوشه پایین و راست صفحه، در زیر تابلوی زرد رنگ Jump to ContigView است(شکل ۲۶-۳). برای این کار مراحل زیر را دنبال کنید:

أ- گزینه Gene را ازمنوی From انتخاب کنید.

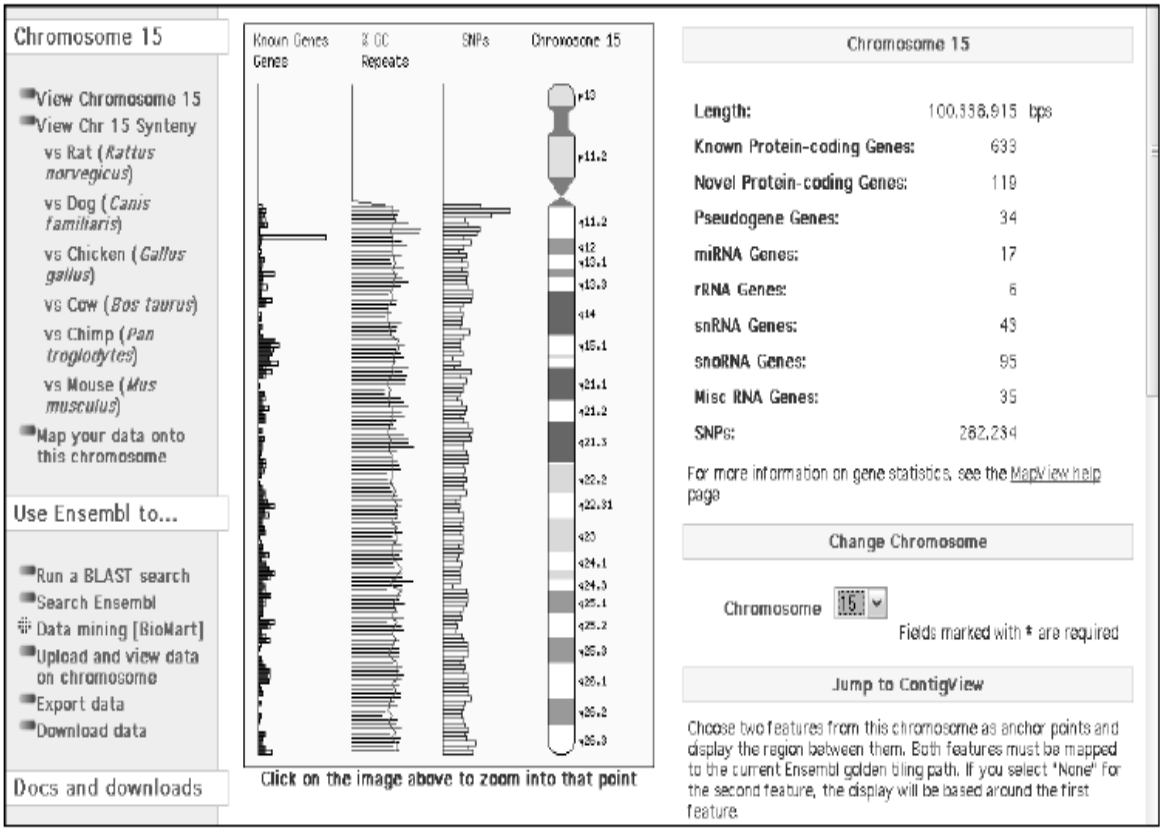
ب- در پنجره جستجو، DUT را وارد کرده و سپس بر روی دکمه Go که به رنگ قرمز است کلیک نمایید. اکنون شما به شکلی پیچیده حاوی اطلاعات مرتبط با ساختار و محتویات ژنومی ژن DUT دست یافته اید، که از بالا به پایین وضوح آن افزایش می یابد: یک شکل شماتیک کلی (در مقیاس 1Mbp ، شکل ۲۷-۳)، شکلی با جزئیات بیشتر (در مقیاس 10 Kbp، شکل ۲۸-۳) و حتی شکلی از جفت باز ها (در مقیاس 100 bp). انتقال نشانگر موس به مناطق مختلف به شما اجازه بررسی بیشتر موقعیت های کروموزومی و

مطالعه ساختار داخلی و شکل های دیگر این ژن خاص و نیـز ارزیـابی تـوالی هـای احتمـالی حـاوی پلـی مورفیسم های تک نوکلئوتیدی[1] (SNP) را می دهد.

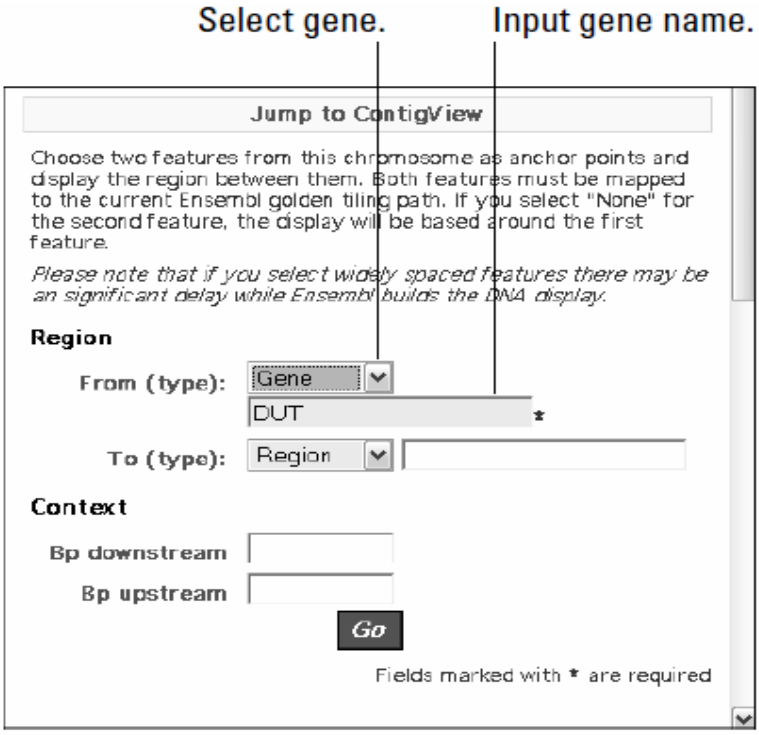


شکل ۲۴-۳: آغاز سفر به ژنوم انسانی

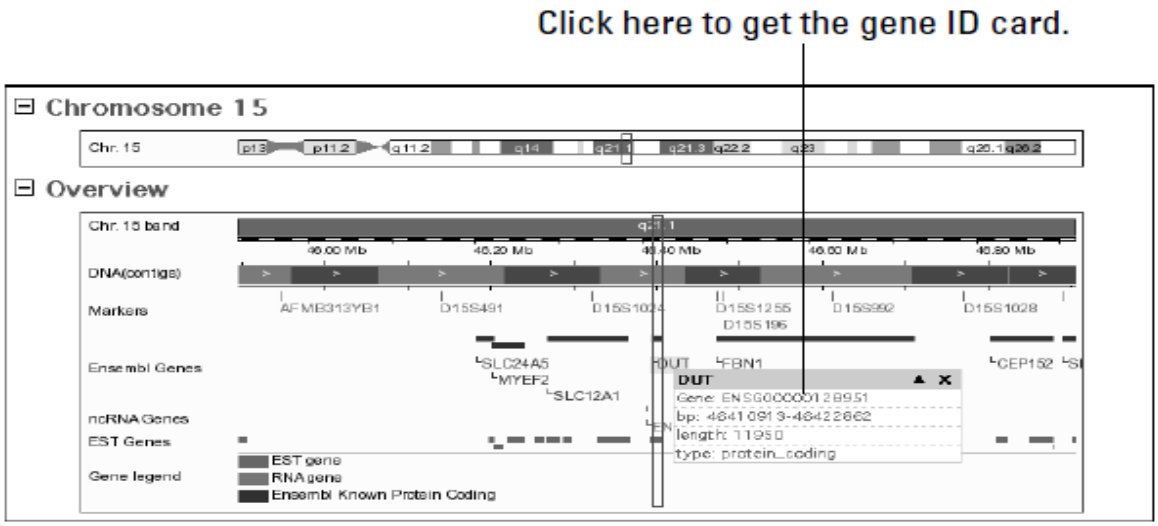
[1] single-nucleotide polymorphisms



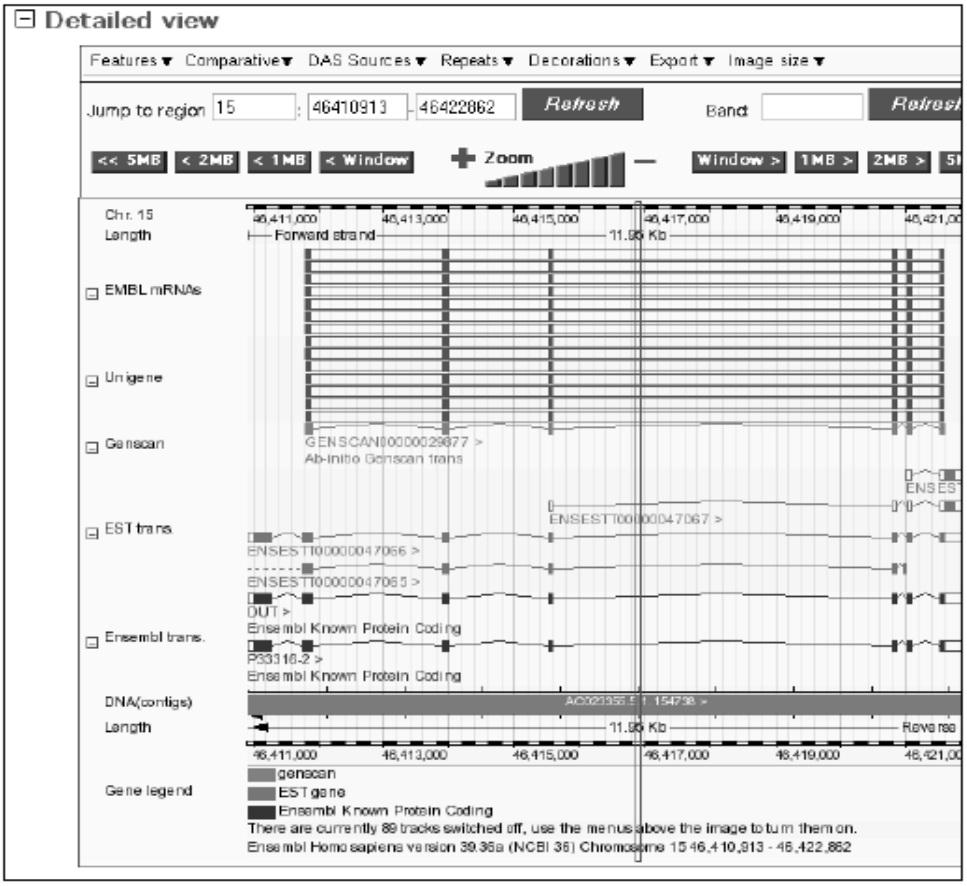
شکل ۲۵-۳: تمرکز بر یک زیر مجموعه از یک ژنوم: کروموزوم ۱۵ انسانی



شکل ۲۶-۳: جستجوی ژن dUTPAse انسانی : DUT



شکل ۲۷-۳: موقعیت ژن انسانی DUT : در مقیاس 1-Mbp



شکل ۲۸-۳: موقعیت ژن انسانی DUT : در مقیاس 10-kb با جزئیات بیشتر

دستیابی به کارت شناسایی ژن DUT انسانی

اگر به شکل ۲۷-۳ مراجعه کنید در آنجا به نام ژن DUT بصورت مشخص برخورد خواهید نمود. در صورت کلیک بر روی نام ژن یک برچسب در مقابل شما ظاهر خواهد شد، که حاوی اطلاعاتی چون شماره ژن در Ensembl، موقعیت آن،

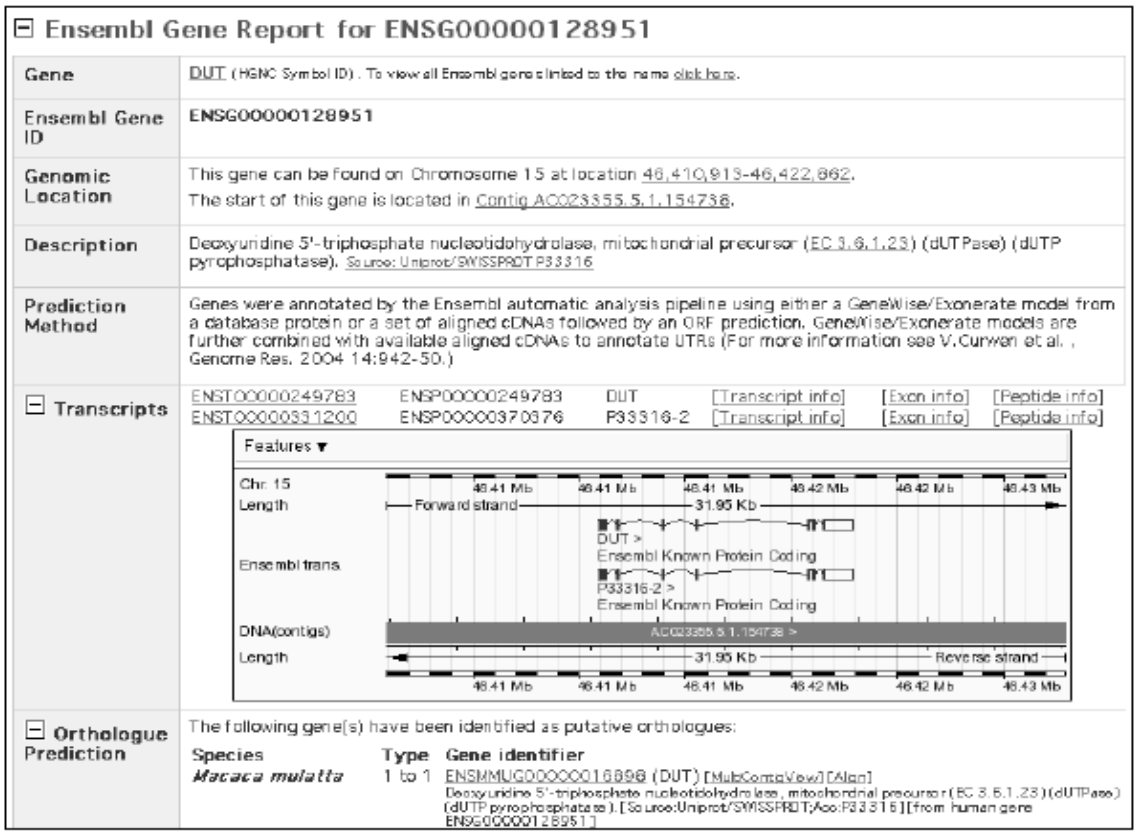
طول آن و نیز نوع آن (کدکننده پروتئین) می باشد. در صورت کلیک بر روی شماره ژن، یک کارت شناسایی جامع شامل تمامی جزئیات مورد نیاز شما در مورد ژن مورد نظر ظاهر می گردد. این اطلاعات یا مستقیما ذکر شده اند یا توسط لینک هایی امکان دسترسی به آن ها در بانک های اطلاعاتی دیگر فراهم شده است. در شکل ۲۹-۳ می توانید تصویر کارت شناسایی ژن DUT انسانی را مشاهده نمایید.

یافتن ژن های بیماری با SNP های کد کننده با استفده از BioMart

اطلاعات بسیار زیادی در Ensembl وجود دارد که می تواند پاسخ بسیاری از پرسش های شما را به آسانی و به صراحت و البته با سرعت بیان نماید. البته ممکن است نحوه کار همیشه به زیبایی آنچه که در فصل های قبل به شما نشان دادیم، نباشد. این سیستم استخراج اطلاعات که کارکنان Ensembl آن را طراحی نموده اند، BioMart می گویند (شکل ۲۳-۳، گوشه سمت چپ و بالا). تصور کنید شما بعنوان یک متخصص ژنتیک انسانی سؤالات زیر را از خود می پرسید:

کدام یک از ژن های کروموزوم ۱۵ با یک بیماری مرتبط هستند؟

کدام پلی مورفیسم های تک نوکلئوتیدی (SNPs) شناسایی شده اند؟



شکل ۲۹-۳: بخش فوقانی کارت شناسایی Ensembl مربوط به ژن انسانی DUT

سؤال اول زمانی مطرح است که بخواهید یک ژن مرتبط با یک بیماری خاص را که روی کر وموزوم ۱۵ واقع شده است پیدا کنید. از طرفی زمانیکه SNP های کد کننده را جستجو می کنید، سؤال دوم مفهوم پیدا می کند. (SNP ها شامل تغییرات نوکلئوتید ی هستند که توالی و حتی شکل و یا عملکرد پروتئین مربوطه را تغییر می دهند.)

شما می توانید به این سؤالات در کمتر از یک دقیقه از طریق Ensembl پاسخ دهید! برای این کار مراحل زیر را دنبال کنید:

۱. به آدرس اینترنتی www.ensembl.org/ مراجعه نمایید. صفحه اصلی Ensembl Project ظاهر می گردد.
۲. بر روی لینک Data mining در گوشه سمت چپ و بالای صفحه کلیک نمایید (شکل ۲۳-۳). صفحه آغاز کننده با عنوان Mart view ظاهر می شود. این صفحه به منظور کار با به روز ترین اطلاعات ژنوم انسانی طراحی شده است (اگر اینگونه نبود dataset را انتخاب کنید).
۳. برای ادامه کار بر روی گزینه Next کلیک نمایید. صفحه ای به ظاهر پیچیده در مقابل شما قرار می گیرد که حاوی تعداد خیلی زیادی گزینه جستجو می باشد. نترسید !! با آنها آشنا خواهید شد.
۴. در قسمت فوقانی صفحه (Regio setup)، مربع کنار گزینه Chromosome را تیک بزنید و از منوی ظاهر شده عدد ۱۵ را انتخاب نمایید.
۵. در بخش بعدی(Gene setup)، مربع بالایی را تیک بزنید، و گزینهWith Disease Association را از منیوی ضمیمه شده انتخاب نمایید. سپس مربع کنار گزینه Gene Type را تیک بزنید و گزینه protein-coding را از منوی مرتبط با آن انتخاب نمایید.
۶. صفحه مربوطه را پایین آورده و قسمت مربوط به تنظیمات جستجوی SNP را پیدا نمایید. مربع کنار گزینه Coding را تیک بزنید و بر روی گزینه Next در قسمت پایین صفحه کلیک نمایید.
۷. در صفحه ای که ظاهر می شود، مربع کنار گزینه های Chromosome Name و Band را تیک بزنید. در مرحله بعد کنار Ensembl Gene ID، Description و MIM Disease ID جستجو نمایید.
MIM مخفف عبارت *M*endelian *I*nheritance in *M*an است، و بهترین بانک اطلاعاتی در زمینه ارتباط بیماری-ژنتیک-انسان تلقی می شود.
۸. به قسمت انتهایی صفحه مربوطه رفته و برروی گزینه Export کلیک نمایید. بلافاصله Ensembl جدولی مشابه شکل ۳۰-۳ در اختیارتان قرار می دهد. جستجوی ما در BioMart، تعداد بسیار زیادی ژن مرتبط با بیماری های گوناگون که پلی مورفیسم های منجر به تغییر پروتئین در آنها ثبت شده اند را دراختیار ما قرار می دهد. این ها بهترین کاندیداها برای جستجوهای بیشتر در این زمینه است. با کلیک بر روی Ensembl Gene ID اطلاعات جامع مرتبط با این ژن ها در اختیارتان قرار خواهد گرفت. با استفاده از این اطلاعات می توانید بررسی کنید که آیا عملکرد آنها با علائم بیماری مرتبط است یا خیر.

وسعت وپیچیدگی پرسش هایی که با مراجعه به منابع Ensembl BioMart قابل بررسی است، بسیار زیاد است. بهتر است وقت بیشتری را صرف امتحان این منبع نمایید، هرچند این ابزار مشابه ابزارهایی که در قسمت های قبلی به آن پرداختیم، جالب نیست.

Chromosome Name	Band	Ensembl Gene ID	Description	MIM Disease ID
15	q11.2	ENSG00000182636	Necdin. [Source:Uniprot/SWISSPROT;Acc:Q99608]	176270
15	q12	ENSG00000104044	P protein (Melanocyte-specific transporter protein) (Pink-eyed dilution protein homolog). [Source:Uniprot/SWISSPROT;Acc:Q04671]	203200
15	q12	ENSG00000104044	P protein (Melanocyte-specific transporter protein) (Pink-eyed dilution protein homolog). [Source:Uniprot/SWISSPROT;Acc:Q04671]	203310
15	q13.3	ENSG00000175344	Neuronal acetylcholine receptor protein, alpha-7 subunit precursor. [Source:Uniprot/SWISSPROT;Acc:P36544]	118511
15	q21.1	ENSG00000171766	Glycine amidinotransferase, mitochondrial precursor (EC 2.1.4.1) (L-arginine:glycine amidinotransferase) (Transamidinase) (AT). [Source:Uniprot/SWISSPROT;Acc:P50440]	602360
15	q21.2	ENSG00000137869	Cytochrome P450 19A1 (EC 1.14.14.1) (Aromatase) (CYPXIX) (Estrogen synthetase) (P-450AROM). [Source:Uniprot/SWISSPROT;Acc:P11511]	107910
15	q21.3	ENSG00000103569	Aquaporin-9 (AQP-9) (Small solute channel 1). [Source:Uniprot/SWISSPROT;Acc:O43315]	602914
15	q21.3	ENSG00000166035	Hepatic triacylglycerol lipase precursor (EC 3.1.1.3) (Hepatic lipase) (HL). [Source:Uniprot/SWISSPROT;Acc:P11150]	151670
15	q24.1	ENSG00000139629	Semaphorin-7A precursor (Semaphorin L) (Sema L) (Semaphorin K1) (Sema K1) (John-Milton-Hargen human blood group Ag) (JMH blood group antigen) (CD108 antigen) (CDw108). [Source:Uniprot/SWISSPROT;Acc:O75326]	607961
15	q24.3	ENSG00000140368	Proline-serine-threonine phosphatase-interacting protein 1 (PEST phosphatase-interacting protein 1) (CD2-binding protein 1) (H-PIP). [Source:Uniprot/SWISSPROT;Acc:Q43586]	604416
15	q26.1	ENSG00000103876	Fumarylacetoacetase (EC 3.7.1.2) (Fumarylacetoacetate hydrolase) (Beta-diketonase) (FAA). [Source:Uniprot/SWISSPROT;Acc:P16930]	276700

شکل ۳۰-۳: نتیجه اولین جستجودر BioMart

فصل چهارم

چگونگی استفاده از بانک های اطلاعاتی توالی و پروتئین

در این قسمت می خوانیم :

- ✓ بررسی چگونگی ساخت و تکامل پروتئین
- ✓ بررسی ورودی UniProtKB/Swiss-Prot از ابتدا تا انتها
- ✓ بررسی جزئیات بیشتر پیرامون بیوشیمی پروتئین
- ✓ کشف و پیش بینی عملکرد پروتئین مورد نظر
- ✓ جستجوی منابع اطلاعاتی مفید پیرامون پروتئین

.....

" انسان یک آنزیم کاتالیز کننده برای تبدیل هوش کربنی به هوش سیلیکونی است."

— Gerard Bricogne

این فصل از کتاب به جستجوی اطلاعات در زمینه توالی های پروتئینی می پردازد. بانک های اطلاعاتی پروتئین صرفا به ارائه توالی ها نمی پردازند. در این فصل به شما نشان خواهیم داد که در صورت تسلط بر شبکه لینک های وب می توانید از طریق پروتئین مربوطه به توالی های ژن آن، عملکرد و ساختار سه بعدی آن دست یابید.

هیچگاه اجازه ندهید که گستردگی و گوناگونی اطلاعات شما را گیج نماید. در دنیای توالی نوکلئوتیدی، جائی که تمامی اطلاعات حول محور یک منبع آنهم GenBank می گذرد، بسیاری از لینک های مرتبط با ژنها و پروتئین ها و عملکرد آنها به UniProtKB/Swiss-Prot وابسته می باشند. این سازمان یک منبع مرکزی تفسیر توالی های پروتئین بوده که اطلاعات موجود در آن بصورت مشترک بوسیله موسسه بیوانفورماتیک سوئیس[1] (SIB) و موسسه بیوانفورماتیک اروپا[2] (EBI) جمع آوری شده است. به علاوه در این فصل چگونگی استخراج کوچکترین اطلاعات مورد نیاز را از این بانک اطلاعاتی آموزش خواهیم داد.

[1] Swiss Institute of Bioinformatics
[2] European Bioinformatics Institute

Swiss-Prot: تصور شخصی از دنیای پروتئین

هر چند GenBank و Swiss-Prot دارای اهداف و کاربرد های مشابهی برای زیست شناسان می باشند، ولی فلسفه کلی این دو مرکز یکسان نمی باشد. GenBank به عنوان یک مخزن توالی های اولیه، از یک نقطه نظر تاریخی نسبتاً سخت گیرانه پیروی می کند. در GenBank ، نویسندگان نسبت به محتوای ورودی هایی که به آنها ارتباط دارند، دارای اختیار کامل می باشند. مفسرین در GenBank تنها نسبت به ورودی های Refseq که اخیراً معرفی شده اند، مسئولیت دارند. این افراد با استفاده از تخصص خود به آنالیز مجموعه ورودی های ارائه شده توسط دانشمندان مختلف، می پردازند. با این وجود، ورودی های مربوط به Refseq هیچگاه جایگزین ورودی های اصلی Genbank نمی شوند، بلکه تمامی لایه های اطلاعاتی در کنار هم نگهداری می شوند.

در مقابل، Swiss-Prot یک بانک اطلاعاتی محسوب نمی شود، بلکه به عنوان یک منبع اطلاعاتی تلقی شده که بر اساس نوع نگرش مسئول و مؤسس این سازمان، Amos Bairoch، و با همراهی گروهی از متخصصین هدایت می شود. در نتیجه تنها نسخه حقیقی Swiss-Prot همان نسخه ای است که در کامپیوتر شخصی Amos قرار دارد. تصور شخصی به معنی آن است که Amos Bairoch نیازی به اجازه هیچ کس برای تصحیح و یا ایجاد تغییر در ورودی Swiss-Prot ندارد. کافی است که Amos توسط متخصصین خود و یا از طریق بررسی مطالب و نوشتجات موجود به لزوم ایجاد این گونه تغییرات پی ببرد.

به وضوح مشخص است که ارزش این فلسفه به قابلیت انعطاف آن است. به این ترتیب خطاهای آشکار به سرعت حذف و کشف های داغ! به سرعت به این منبع اضافه می شوند. دقیقاً به همین خاطر است که Swiss-Prot به عنوان بهترین بانک اطلاعاتی آنالیز پروتئین شناخته شده است و نقش مؤثری را در زمینه بیولوژی مولکولی و ژنومیک ایفا می کند. جنبه منفی این قضیه محدودیت ها و مشکلات اجتناب ناپذیری است که هر محققی در دنیا مرتکب می شود، حتی زمانی که توسط یک گروه بزرگ آنالیزور در مؤسسه بیوانفورماتیک اروپا (EBI) در Hinxton و نیز موسسه بیوانفورماتیک سوئیس (SIB) در Geneva و متخصصان دیگر حمایت می شود. در حال حاضر، Swiss-Prot نمی تواند از عهده حجم بالای توالی های نوکلئوتیدهای جدید شناسایی شده برآید، بنابراین از نظر کامل بودن در بین بانک های اطلاعاتی در رتبه های پایین تر قرار می گیرد.

به منظور رفع این قبیل مشکلات، یک حافظه موقت در Swiss Prot ابداع گردیدکه به آن TrEMBL[1] گفته می شود. ورودی های TrEMBL در EBI از مطالب ارائه شده به GenBank ساخته می شوند، و اغلب به صورت اتوماتیک و با معیار تشابه توالی ها، آنالیز و تفسیر می شوند. در نهایت با بازرسی بصری، تصحیح دستی و تأیید نهایی توسط Amos Bairoch، ورودی های TrEMBL به ورودی های Swiss Prot تغییر می یابند. شاید این حقیقت که یک منبع اطلاع رسانی با گسترش جهانی و نقشی کلیدی در جهان توسط یک نفر اداره می شود شما را متعجب ساخته است. اما باید گفت این روز ها اینگونه مسائل در بیوانفورماتیک بسیار به چشم می خورد. از مثال های مشهور در این زمینه موارد زیر را می توان نام برد:

بانک اطلاعاتی توالی ایمنوگلوبولین[2] مربوط به Elvin Kabat و یا بانک اطلاعاتی آنزیم های محدودکننده[3] مربوط به Richard J.Robert و همچنین بانک اطلاعاتی بیماری های ژنتیکی انسان که توسط Victor McKusick اداره می شود (توارث مندلی در انسان بصورت آنلاین[4] (OMIM)).

به قول Winston Churchill که می گوید:

در بیوانفورماتیک، به ندرت "خیلی زیاد" چیزی به "خیلی کم" قرض داده است!

[1] automatic *TR*anslation of *E*uropean *M*olecular *B*iology *L*aboratory nucleotide sequences
[2] Immunoglobulin sequence database
[3] Restriction Enzyme Database
[4] Online Mendelian Inheritance in Man (OMIM)

راههای بسیاری برای استفاده از بانک اطلاعاتی Swiss-Prot (که امروزه به آن پایگاه علمی UniProt می گویند) وجود دارد. بسیاری از پایگاه های ژنومیک، جستجوگر های ژنوم و یا سایت هایی که حاوی یک سیستم بازیابی توالی[1] (SRS) هستند، شما را بطور مستقیم با ورودی Swiss-Prot مرتبط می کنند. اگر می خواهید بدانید چگونه در Swiss-Prot یک ورودی را از طریق بکارگیری کلید واژه جستجو کنید، به فصل ۲ مراجعه کنید.

این فصل به شما در درک بهتر محتویات هر ورودی Swiss-Prot کمک می کند، تمام قسمت های موجود در این بانک اطلاعاتی را بشناسید. سپس نحوه استفاده از این تفاسیر و آنالیز ها را به عنوان نقطه شروعی برای دسترسی به گستره ای از بانک اطلاعاتی تخصصی به شما نشان داده می شود. شما می توانید از این بانک های اطلاعاتی برای بررسی بیشتر عملکرد و ساختار پروتئین مورد نظرتان استفاده نمایید.

این بخش را با توضیح چگونگی تبدیل توالی پروتئین ابتدائی به یک پروتئین بالغ توسط سلول، شروع می کنیم. تغییرات پس از ترجمه نقش بسازیی در عملکرد پروتئین ایفا می کند. یکی از دلایل استفاده از بانک اطلاعاتی پروتئین، بررسی این قبیل تغییرات می باشد(درصورت تمایل به یادگیری چگونگی پیش گویی این تغییرات به فصل ۶ مراجعه نمایید).

از **ORF** های ترجمه شده تا پروتئین کامل

از آنجائیکه توالی یابی مولکول های DNA بسیار آسانتر و ساده تر از پروتئین ها است، بسیاری از توالی های آمینواسیدی شناسایی شده در حقیقت با ترجمه کامپیوتری ORF ها که از طریق آنالیز اطلاعات ژنومیک شناسایی شده اند، به دست آمده اند. بسیاری از این پروتئین ها در عمل هیچگاه جداسازی و شناسایی نشده اند. به همین علت زمانیکه متخصصین بیولوژی مولکولی و بیوانفورماتیک در مورد ORF های ترجمه شده مثل پروتئین های حقیقی صحبت می کنند، متخصصین پروتئین نسبت به ما احساس تنفر می کنند! در آغاز کار ژنومیک، ساده ترین راه برای ما این بود که متخصصین پروتئین را به عنوان انسان هایی بد اخلاق و قدیمی پنداشته و آنها را نادیده بگیریم. با این وجود، پس از توالی یابی ژنوم، توجه دانشمندان زمینه پروتئین به بررسی پروتئین ها در مقیاس بزرگ، که به آن پروتئومیک[2] گفته می شود، معطوف شد. پروتئومیک رشته ای علمی است که به تجسم و تصویر سازی گروهی از مولکول های پروتئینی موجود در یک بافت و یا ارگانیسم خاص می پردازد. آنالیز پروتئومیک حجم گسترده ای از اطلاعات را به ما ارائه می دهد و به سادگی نشان می دهدکه متخصصین پروتئین دلایل خوبی برای عصبانیت خود داشته اند.

ORF ها: همواره آن چیزی که می بینی همان چیزی نیست که به دست می آوری!

هنگامیکه زیست شناسان ژل الکتروفورز دو بعدی را تهیه نمودند(یک متد جالب آزمایشگاهی که مولکول های پروتئین را در یک جهت بر اساس وزن و در جهت دیگر بر اساس بار الکتریکی جدا می کند.)، مشاهده کردندکه پروتئین های حقیقی رفتاری مشابه آنچه در ترجمه کامپیوتری بدست آمده بود، از خود نشان نمی دادند. در دنیای ORF و پروتئین ها، همواره آن چیزی که می بینی همان چیزی نیست که به دست می آوری.

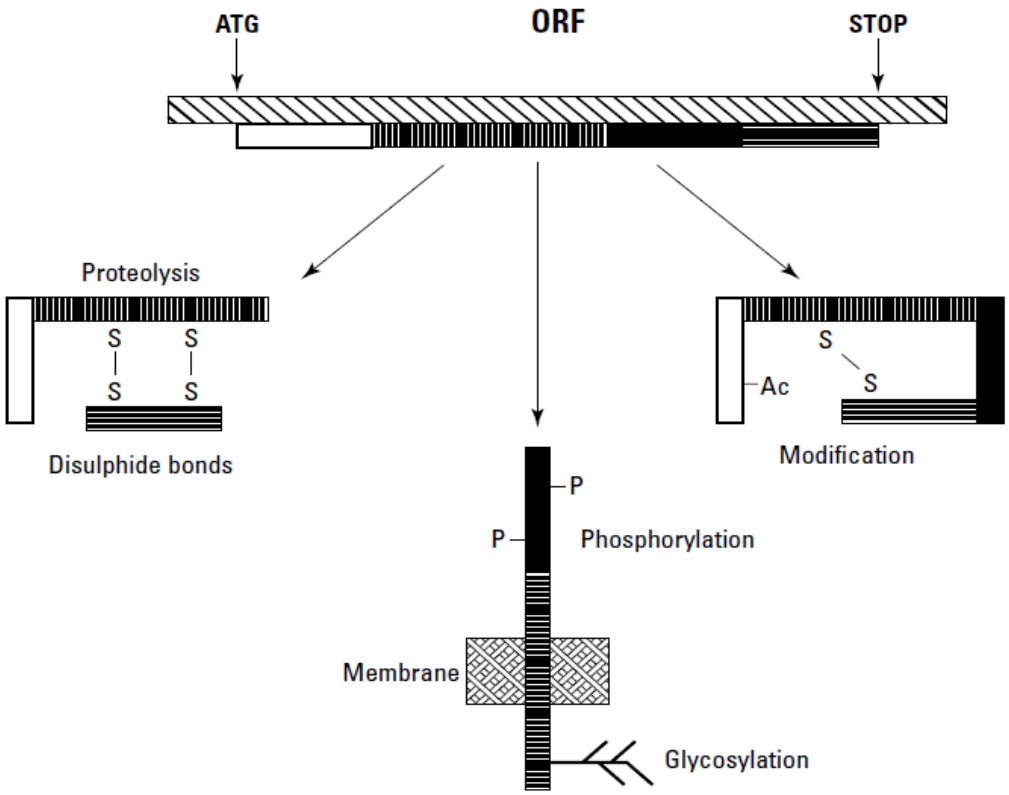
[1] sequence retrieval system
[2] proteomics

زمانیکه پروتئین از روی RNA ترجمه میگردد، زنجیره آمینواسیدی تشکیل شده، طی تبدیل به پروتئین بالغ، متحمل تغییرات گسترده ای می شود. حتی پیش گویی خصوصیات فیزیکوشیمیایی پروتئین نهایی (مانند اندازه، وزن مولکولی و یا نقطه ایزوالکتریک) از روی توالی آمینواسیدی شبیه سازی شده توسط کامپیوتر، کاری سختی است. به این پروسه پیچیده در تکامل پروتئین، تغییرات پس از ترجمه[1] گفته می شود، که در شکل ۱-۴ به صورت خلاصه نمایش داده شده است.

بطور کلی بالغ شدن پروتئین شامل ترکیبی از مراحل زیر می باشد:

- ✓ برش هایی درون زنجیره آمینواسیدی
- ✓ جداسازی قطعه هایی از زنجیره آمینواسیدی (به عنوان مثال در انسولین)
- ✓ تغییرات شیمیایی در بعضی از آمینواسید ها (به عنوان مثال متیله شدن[2])
- ✓ اضافه شدن مولکول های لیپیدی (به عنوان مثال پدیده myristoylation)
- ✓ اضافه شدن مولکول های گلیکوزیدی (گلیکوزیله شدن[3])

از جمله نقش های مهم یک بانک اطلاعاتی پروتئین (در مقایسه با مجموعه توالی های ORF) ارائه چنین اطلاعاتی می باشدکه اغلب توسط داده های تجربی و یا پیش گویی با استفاده از تکنیک های خاص پردازش اطلاعات بدست آمده اند (برای کسب اطلاعات بیشتر به فصل ۶ مراجعه کنید). از این رو تغییرات پس از ترجمه، بخش قابل توجهی از ویژگی های پروتئین موجود در ورودی های Swiss-Prot را تشکیل می دهد.



شکل ۱-۴: تغییرات احتمالی در حین بالغ شدن پروتئین

[1] post-translational modification
[2] methylation
[3] glycosylation

مقصد نهایی هر پروتئین

عملکرد صحیح هر پروتئین نیازمند اینست که به مقصد و هدفی صحیح در یک ارگانیسم و یا در یک سلول برسد. پس از ترجمه، زنجیره پپتیدی ممکن است با سیگنال های مختلفی که اختصاصی توالی هستند، مواجه شود (ZIP codes). سلول از این سیگنال ها برای هدایت پروتئین به محل مناسب (در داخل و یا خارج از سلول) استفاده می کند. این دسته بندی پروتئین ها، مستلزم عبور پروتئین از یک یا چندین غشاء می باشد و به اصطلاح به این نوع انتقال Translocation می گویند. مقاصد و فعالیت های نهایی هر پروتئین شامل موارد زیر می شود:

- ✓ اتصال به غشاء سلولی
- ✓ ترشح به خارج از سلول
- ✓ انتقال به فضای پری پلاسمی (در مورد باکتری ها)
- ✓ انتقال به میتوکندری و سایر اندامک ها
- ✓ انتقال به هسته سلول

از آنجائیکه مقصد نهایی هر پروتئین تعیین کننده عملکرد آن است لذا این اطلاعات، به صورت اثبات شده یا پیش بینی شده، یکی از مهمترین شاخص های موجود در بانک اطلاعاتی پروتئین نظیر Swiss-Prot می باشند.

گوناگونی و تنوع ترکیبی عملکرد و نحوه تا خوردن

مهمترین مرحله تبدیل پروتئین سنتز شده به یک پروتئین کارآمد، تا خوردن رشته و ایجاد ساختاری سه بعدی پایدار و فشرده می باشد. به جز پروتئین های کوچک (کوچکتر از ۱۰۰ اسید آمینه)، ساختار نهایی هر پروتئین بطور معمول شامل چندین دامین نسبتاً مستقل است. شما می توانید این دامین ها را به عنوان قطعات پازل انعطاف ناپذیر و محکمی تصور نمایید. طبیعت از این قطعات پازل، طیف گسترده ای از پروتئین ها را می سازد. برای مثال با داشتن سه دامین به نام های A ، B و C، قادر هستید پروتئین هایی به صورت های زیر ایجاد نمایید:

AAA، ABC، BCC، BAC و...

اغلب پروتئین های طبیعی، ترکیبی از ۱ تا ۱۰ دامین هستندکه از میان چند هزار دامین انتخاب شده اند.. با وجود گوناگونی زیاد توالی ها، این دامین ها بوسیله *scaffold sequence signatures* (واحد های تکراری در محتوای آمینواسیدی پروتئین ها که با وجود میلیون ها سال تکامل واگرا[1] هم قابل شناسایی هستند)، شناسایی می شوند. شناسایی و توصیف دامین های پروتئینی یکی از مهمترین و گسترده ترین موضوعات قابل بررسی در بیوانفورماتیک می باشد(به فصل ۶، ۷، ۹ و ۱۱ مراجعه نمایید). دانستن معماری نحوه قرار گرفتن دامین های یک توالی خاص پروتئین بسیار حائز اهمیت است ، زیرا می تواند حاوی نکاتی در مورد ساختار سه بعدی احتمالی پروتئین ، و نیز عملکرد سلولی و بیوشمیایی آن باشد. به همین دلیل از مهمترین قسمت های هر ورودی Swiss-Prot ، توضیحاتی پیرامون سازماندهی دامین ها در پروتئین می باشد، که با استفاده از متد های موجود شناسایی و پیش بینی شده است.

[1] divergent

مطالعه ی یک ورودی Swiss-Prot

با وجود مفاهیم پیچیده ای که در توصیف یک پروتئین به چشم می خورد در حقیقت پروتئین ها نسبت به ژن ها ساده تر هستند:

- ✓ پروتئین ها توالی های نسبتاً کوچکی دارند(بطور متوسط ۳۵۰ آمینواسید)
- ✓ بر خلاف ژن ها، پروتئین ها دارای ابتدا و انتهای مشخصی هستند.
- ✓ پروتئین ها به صورت تک رشته ای می باشند.
- ✓ با وجود تغییراتی که طی تبدیل توالی ORFs به پروتئین نهایی صورت می گیرد، ترتیب پشت سر هم قرار گرفتن آمینواسید ها ثابت می ماند.

به استناد همه دلایل ذکر شده، احتمالاً انتظار دارید که خواندن و فهم ورودی های Swiss-Prot بسیار ساده تر از ورودی های GenBank باشد. انتظار شما درست است.

مشابه یک ورودی GenBank (که در فصل ۳ به آن پرداخته شد)، یک ورودی Swiss-Prot به چند بخش عمده تقسیم می گردد:

- بخش اطلاعات عمومی
- بخش اطلاعات مربوط به منابع مورد استفاده[1]
- بخش اطلاعات عملکردی
- جدول شاخص ها
- بخش توالی

به عنوان مثال در قسمت های بعدی بخش های مختلف ورودی Swiss-Prot مربوط به رسپتور انسانی فاکتور رشد اپی درمال (EGFR) را برای شما از ابتدا تا انتها و مرحله به مرحله نشان خواهیم داد.

کشف رموز ورودی EGFR در Swiss-Prot

رسپتور EGFR یک مولکول نسبتاً پیچیده است. این رسپتور را انتخاب نمودیم زیرا به وضوح ارزش اطلاعاتی را که می توانید از Swiss-Prot به دست بیاورید را به شما نشان می دهد. فصل ۲، راه های ماهرانه ای را برای جستجو در Swiss-Prot به شما نشان می دهد، ولی اکنون به منظور رسیدن به اهدافمان بهتر است مطابق دستوالعمل زیر جستجوی خود را انجام دهیم:

۱. به آدرس اینترنتی www.expasy.ch/sprot/ مراجعه نمایید.
صفحه مربوط به UniProtKB/Swiss-Prot از سرور ExPASy ظاهر می گردد (شکل ۱۰-۲، فصل ۲).
۲. در پنجره جستجو(در قسمت بالایی صفحه) کد شناسایی P00533 را تایپ نمایید.

این ورودی مربوط به رسپتور انسانی فاکتور رشد اپی درمال می باشد، و چون یک پروتئین داغ! می باشد، آن را انتخاب کرده ایم. این پروتئین یک انکوژن[2] (مستعد کننده ابتلا به سرطان) می باشد و جزئیات کامل و خوبی را در

[1] bibliographic informations
[2] oncogene

اختیارمان قرار می دهد. از طرفی چون اکثر دانشمندان روی این قبیل پروتئین های داغ! کار می کنند، تفاسیر و نتایج بیشتری در مورد این گونه پروتئین ها وجود دارد.

۳. بر روی گزینه Go کلیک نمایید.

صفحه مربوط به ورودی UniProtKB/Swiss-Prot P00533 ظاهر می گردد. در صورت تمایل می توانید با کلیک بر روی گزینه Printer-Friendly View، در بالای سمت راست صفحه، این صفحه را پرینت نمایید.

به علت گستردگی اطلاعات موجود پیرامون این پروتئین خاص، ممکن است این صفحه پیچیده ترین صفحه ای باشد که در Swiss-Prot با آن مواجه می شوید. اغلب ورودی های دیگر واژگان کلیدی و فیلد هایی را که در این مکان می یابید، نشان نخواهند داد.

اطلاعات کلی پیرامون ورودی مورد نظر

اکنون آماده اید تا مرحله به مرحله از میان این ورودی پیچیده ولی مفید عبور نمایید. خطوط آبی رنگ قسمت های اصلی و مهم را از یکدیگر جدا می نماید. اولین قسمت با عنوان Entry Information آغاز می گردد و شامل شش فیلد زیر می باشد:

- ✓ **نام ورودی: *EGFR_HUMAN*.** این قسمت از Swiss-Prot، ایده ای کلی از ورودی مربوطه را ارائه می دهد. این نام واضح و صریح به نظر می رسد ولی به آن زیاد اعتماد نکنید. این وضوح همواره آنچیزی نیست که مدنظر است! برای مثال ، نام ورودی مربوط به fly lysozyme، *LYSA_DROME* می باشد. یقیناً به خاطر سپردن آن مشکل است (DROME اشاره به *DRO*sophila *ME*lanogaster دارد).

به عنوان یک قانون کلی به یاد داشته باشید که نام یک ورودی همواره معنا دار و پایدار نیست. در نتیجه اکیداً به شما پیشنهاد می کنیم از آن برای جستجو در منابع اینترنتی دیگر، و یا به عنوان نام توالی در یک مقاله استفاده نکنید. در عوض، می توانید از شماره دسترسی اولیه[1] که در بخش با آن آشنا می شوید، استفاده نمایید.

- ✓ **شماره دسترسی اولیه: P00533.** این شماره یک شناساگر ثابت و بدون تغییر و البته منحصر به فرد برای جستجوی توالی مربوطه می باشد. هنگامی که می خواهید یک ورودی را در منابع یک مقاله ذکر کنید، باید از شماره دسترسی استفاده نمایید. شماره های دسترسی بسیار مفیدند و شما بوسیله آنها می توانید تاریخچه یک پروتئین در یک بانک اطلاعاتی را بدست آورید. برای مثال، تصور کنید پروتئین EGFR (یک پروتئین خیلی بلند) در ابتدا بصورت بخشی از یک توالی معرفی شد. سپس دانشمندان با کار بیشتر قسمت های بلندتری از این پروتئین را شناسایی کردند. در هر مرحله، آنها نسخه جدیدتری از پروتئین رسپتور EGFR را منتشر و ارائه کردند (شاید در پروسه تصحیح اشتباهات قبلی). برخلاف GenBank، Swiss-Prot فقط یک نسخه از رسپتور EGF را نگهداری کرده و اطلاعات بدست آمده تا به امروز را خلاصه می نماید. نکته دیگر آنکه، شماره های دسترسی کهنه و قدیمی هیچگاه به کلی کنار گذاشته نشده اند، بلکه در فیلد شماره دسترسی ثانویه[2] ذخیره می گردند.
- ✓ **شماره دسترسی ثانویه: *Q14225, P06268,* O00688 و ....** شماره های دسترسی ثانویه این اطمینان را به شما می دهند که هنگام مطالعه یک مقاله قدیمی و مشاهده Q12225 و یا P0 2 8 به عنوان شماره های دسترسی می

[1] primary accession number
[2] secondary accession number

توانید از این شماره های دسترسی قدیمی برای دستیابی به جدید ترین نسخه EGFR استفاده نمایید. سعی کنید یکی از این شماره های قدیمی را برای جستجو استفاده کنید. با تایپ هر یک از شماره های دسترسی ثانویه، به شماره دسترسی اصلی که همان P00533 می باشد ارجاع داده خواهید شد. این یک راه ساده و روشن برای حذف ورودی های قدیمی از طریق نگهداری آنها به صورت لینکهایی به مسیر اصلی است. قسمت های بعدی در بخش "اطلاعات کلی"، نیاز به توضیح ندارد:

- ✓ **تاریخ ورود به Swiss-Prot: *July 21, 1986***
- ✓ **تاریخ آخرین تغییرات اعمال شده در توالی: *November 1, 1997*.**

این مطلب نشانگر آن است که دستیابی به توالی کامل و درست در حدود ۱۰ سال طول کشیده است.

- ✓ **تاریخ آخرین تغییرات اعمال شده در تفاسیر توالی: *July 25, 2006***

البته واضح است که هنوز کارهای بسیار زیادی بر روی این پروتئین مهم در حال انجام است.

توجه کنیدکه در هر ورودی، فیلد definition (قسمت چپ هر بخش) یک لینک فعال است که شما را به هر بخش از ورودی که بخواهید راهنمایی می کند. با مطالعه بخشی از آن حتماً متخصصین بیوانفورماتیک را تحسین خواهید نمود.

نام و منشأ یک پروتئین

در این قسمت، با بخش جدیدی (که به صورت یک تابلوی آبی مشخص شده است) مواجه می شوید: Name and Origin of the Protein. این بخش شامل ۵ فیلد می باشد:

- ✓ **Protein Name: *Epidermal growth factor receptor [Precursor]*** این فیلد شامل توصیفی کلی از توالی مربوطه می باشد و برای شناسایی دقیق پروتئین، بسیار حرفه ای و کامل طراحی شده است. در صورتیکه توالی مربوطه فقط یک قطعه آمینواسیدی است، این فیلد آنرا نشان می دهد.

  در این قسمت [Precursor] بدین معنی است که توالی مربوطه باید شکسته شود تا پروتئین بالغ بدست آید.
- ✓ **Synonyms: *Receptor tyrosine-protein kinase ErbB-1, EC 2.7.10.1*** شماره EC (2.7.10.1) نمایانگر واکنش بیوشمیایی (tyrosine-protein kinase) است که این پروتئین انجام می دهد. EC مخفف Enzyme Nomenclature Committee است و در رابطه با آن در جدول ۱-۴ اطلاعات بیشتری آورده شده است. این قسمت خیلی هم گیرا و مؤثر به نظر نمی رسد، ولی می تواند نقطه شروعی برای یک سفر سحرانگیز و جذاب برای درک کامل عملکرد آنزیمی این پروتئین باشد. برای مثال، بر روی لینک 2.7.10.1 کلیک نمایید تا شمای NiceZyme این آنزیم را مطابق آنچه در شکل ۲-۴ مشاهده می نمایید، ببینید. لینک های موجود در این صفحه انواع متفاوتی از اطلاعات را در اختیارتان می گذارند. در این قسمت شما می توانید:
  - پروتئین خود را درون چارت مسیر های متابولیک سلولی مکان یابی نمایید،
  - به "توضیحات" و "پروفایل توالی" خانواده ای که توالی به آن تعلق دارد دست یابید،
  - به فرمول های شیمیایی مواد بیوشیمیایی که پروتئین مربوطه با آن واکنش می دهد دسترسی پیدا نمایید.
- ✓ **Gene Name: *EGFR*** این فیلد زمانیکه با بانک های اطلاعاتی توالی نوکلئوتیدی و ژنوم سر و کار دارید می تواند مفید باشد. در واقع کمک می کند که جستجوهای شفاف و بدون ابهامی داشته باشید(در صورت تمایل به بررسی بیشتر نام های ژن ها به فصل ۳ مراجعه نمایید).

- ✓ **From**: ***Homo sapiens (Human) [TaxID:9606]*** فیلد From، منشأ پروتئین مربوطه را به دقت مشخص می کند. در اغلب موارد معرفی گونه شامل نام گونه و جنس به صورت لاتین و سپس نام انگلیسی که در پرانتز قرار می گیرد. برای ویروس ها، تنها نام انگلیسی رایج آنها آورده می شود.با کلیک بر روی لینک Toxonomic ID 9606، وارد صفحه *H. sapiens* در بانک اطلاعاتی تاکسونومی NEWT که توسط گروه UniProtKB پشتیبانی می شود(www.ebi.ac.uk/newt/index.html)، می شوید.
- ✓ **Toxonomy**: ***Eukaryota; Metazoa; Chordata; Craniata; Vertebrata و ...:*** این فیلد لیستی از دسته بندی تاکسونومیک ارگانیسم های منبع را در اختیارتان می گذارد. در این قسمت از دسته بندی تاکسونومیکی استفاده می شود که در مرکز ملی اطلاعات بیوتکنولوژی (NCBI) نگهداری می شود. نام هرگروه یک لینک برای دسترسی به لیست ورودی های Swiss-Prot مربوط به آن گروه می باشد.

این فیلد راهی آسان برای جمع آوری تمامی توالی های پروتئین یک گونه در پروژه های تحقیقاتی می باشد.

منابع

بخش بعدی در ورودی P00533، بخش منابع است و همانطور که به نظر می رسد بسیار ساده و راحت است. لیستی از تمامی منابع و کتب استفاده شده در ساخت ورودی مربوطه در این بخش به چشم می خورد. در مورد مثال ما، این بخش شامل مطالعات گوناگونی پیرامون توالی یابی، توصیف ایزوفرم[1] های پروتئین مربوطه که در بافت ها و تومورهای گوناگون بیان شده اند و همچنین نقشه تغییرات پس از ترجمه، باند های دی سولفیدی و محل های اتصال لیگاند، می شود. اهمیت این پروتئین بوسیله تعداد منابع ذکر قابل سنجش است؛ ۳۴ منبع پراهمیت بودن آن را نشان می دهد. Swiss Prot شماره شناسایی NCBI/PubMed هریک از مقالات ذکر شده در بخش منابع، و در اکثر موارد لینک مستقیم به خلاصه مقاله را ارائه می کند.

NiceZyme View of ENZYME: EC 2.7.10.1

Official Name
Receptor protein-tyrosine kinase.
Alternative Name(s)
Receptor protein tyrosine kinase.
Reaction catalysed
ATP + a [protein]-L-tyrosine <=> ADP + a [protein]-L-tyrosine phosphate
Comment(s)
• The receptor protein-tyrosine kinases, which can be defined as having a transmembrane domain, are a large and diverse multigene family found only in metazoans. • In the human genome, 58 receptor-type protein-tyrosine kinases have been identified and these are distributed into 20 subfamilies. • Formerly EC 2.7.1.112.
Cross-references

PROSITE	PDOC00100
BRENDA	2.7.10.1
PUMA2	2.7.10.1
PRIAM enzyme-specific profiles	2.7.10.1
Kyoto University LIGAND chemical database	2.7.10.1
IUBMB Enzyme Nomenclature	2.7.10.1
IntEnz	2.7.10.1
MEDLINE	Find literature relating to 2.7.10.1
MetaCyc	2.7.10.1

P13368, 7LES3_DROME; P20806, 7LES3_DROVI; Q5UN73, ALK_HUMAN;
P97793, ALK_MOUSE; Q9VBW3, CAD96_DROME; P18460, CEK2_CHICK;
P16461, CEK3_CHICK; P13369, CSF1R_FELCA; P07333, CSF1R_HUMAN;
P09581, CSF1R_MOUSE; Q00495, CSF1R_RAT; Q08345, DDR1_HUMAN;

شکل ۲-۴: ورودی Swiss-Prot P00533 به فرم NiceZyme

[1] isoform

بخش Comments

این بخش شامل اطلاعات مفیدی است که در بخش های دیگر نمی توانستند قرار بگیرند. مطالب این قسمت به صورت متون دسته بندی شده در قالب بلوک هایی می باشد. یک بلوک شامل یک یا چند خط می باشد که توسط یک کلید واژه اصلی، مشخص می شود. کلید واژه های اصلی خود رنج گسترده ای را شامل می شود. این مطلب در شکل ۳-۴ به وضوح نشان داده شده است. برای دسترسی به این لیست، که بخشی از دفترچه راهنمای Swiss-Prot را تشکیل می دهد، می توانید بر روی سر فصل Comments کلیک نمایید).

Topic	Description
ALLERGEN	Information relevant to allergenic proteins
ALTERNATIVE PRODUCTS	Description of the existence of related protein sequence(s) produced by alternative splicing of the same gene, alternative promoter usage, ribosomal frameshifting or by the use of alternative initiation codons; see 3.21.15
BIOPHYSICOCHEMICAL PROPERTIES	Description of the information relevant to biophysical and physicochemical data and information on pH dependence, temperature dependence, kinetic parameters, redox potentials, and maximal absorption; see 3.21.8
BIOTECHNOLOGY	Description of the use of a specific protein in a biotechnological process
CATALYTIC ACTIVITY	Description of the reaction(s) catalyzed by an enzyme [1]
CAUTION	Warning about possible errors and/or grounds for confusion
COFACTOR	Description of any non-protein substance required by an enzyme for its catalytic activity
DEVELOPMENTAL STAGE	Description of the developmentally-specific expression of mRNA or protein
DISEASE	Description of the disease(s) associated with a deficiency of a protein
DOMAIN	Description of the domain structure of a protein
ENZYME REGULATION	Description of an enzyme regulatory mechanism
FUNCTION	General description of the function(s) of a protein
INDUCTION	Description of the compound(s) or condition(s) that regulate gene expression
INTERACTION	Conveys information relevant to binary protein-protein interaction 3.21.12
MASS SPECTROMETRY	Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods; see 3.21.23
MISCELLANEOUS	Any comment which does not belong to any of the other defined topics
PATHWAY	Description of the metabolic pathway(s) with which a protein is associated
PHARMACEUTICAL	Description of the use of a protein as a pharmaceutical drug
POLYMORPHISM	Description of polymorphism(s)
PTM	Description of any chemical alternation of a polypeptide (proteolytic cleavage, amino acid modifications including crosslinks). This topic complements information given in the feature table or indicates polypeptide modifications for which position-specific data is not available.
RNA EDITING	Description of any type of RNA editing that leads to one or more amino acid changes
SIMILARITY	Description of the similarity(s) (sequence or structural) of a protein with other proteins
SUBCELLULAR LOCATION	Description of the subcellular location of the mature protein
SUBUNIT	Description of the quaternary structure of a protein and any kind of interactions with other proteins or protein complexes, except for receptor-ligand interactions, which are described in the topic FUNCTION.

شکل ۳-۴: کلید واژه های اصلی در بخش Comments بانک اطلاعاتی Swiss-Prot

بخش Comments ورودی P00533 بسیار غنی بوده و حاوی اطلاعاتی با عناوین زیر می باشد:

- ✓ **FUNCTION (عملکرد):** رسپتور EGF مسئول کنترل رشد سلولی است.
- ✓ **CATALYTIC ACTIVITY(فعالیت کاتالیزوری):** تیروزین-پروتئین کیناز
- ✓ **SUBUNIT (زیر واحد ها):** سایر پروتئین هایی که با آن کمپلکس های پایداری را می سازند.
- ✓ **INTERACTION (واکنش ها):** سایر پروتئین هایی که با آن بر هم کنش دارند.
- ✓ **SUBCELLULAR LOCATION (مکان پروتئین در سلول):** پروتئین غشای سلولی
- ✓ **ALTERNATIVE PRODUCTS (محصولات فرعی):** ایزوفرم های حاصل از طریقه های گوناگون اتصال زیر واحد ها
- ✓ **TISSUE SPECIFICITY (اختصاصیت برای یک بافت):** بافت جفت، بافت سرطان تخمدان

- ✓ **PTM:** لیستی از بعضی تغییرات پس از ترجمه
- ✓ **DISEASE (بیماری):** اختلال در EGFR با سرطان ریه مرتبط است.
- ✓ **MISCELLANEOUS (متفرقه):** نشان دهنده مکانیسم مولکولی (دیمر شدن[1]) عملکرد پروتئین می باشد.
- ✓ **SIMILARITY (تشابهات):** تشابه با سایر توالی های زیرخانواده رسپتور EGF
- ✓ **WEB RESOURCE (منبع تحت وب):** لینکی برای اتصال به GeneTests (بانک اطلاعاتی که سرمایه گذار آن NIH[2] است، حاوی آزمون های ژنتیکی بالینی در دسترس)

محتویات موجود در بخش Comments اغلب برای تصمیم گیری در رابطه با انطباق یا عدم انطباق جستجوی شما با یک BLAST یا یک ورودی Swiss-Prot کافی است. اکنون بیایید از بخش Copyright notice بگذریم و به سراغ بخش گسترده ای با عنوان Cross –References برویم.

بخش Cross-Refereces

بخش Cross-References شامل لینک های متعددی برای ورود به ورودی های بانک های اطلاعاتی دیگر می باشد، که اطلاعاتی پیرامون پروتئین مورد نظر ما دارند.

بیشتر لینک های موجود در این بخش شما را به بانک های اطلاعاتی جدیدی راهنمایی می کنند. انتقال از پایگاهی به پایگاه دیگر در وب ممکن است باعث شود در اینترنت گم شوید! و ورودی اصلی Swiss-Prot را از دست بدهید. برای جلوگیری از این اتفاق، لینک های مربوطه را در پنجره ای جدید باز نمایید؛ به این صورت که بر روی لینک مربوطه راست-کلیک نموده و گزینه Open in a New Window را از منوی باز شده انتخاب نمایید.

در بخش Cross-References فیلد های متعددی موجود است. لیست زیر فهرست تعدادی از فیلد های اصلی می باشد:

- ✓ **EMBL:** این فیلد شامل تمامی لینک های ضروری برای اتصال به دنیای توالی های نوکلئوتیدی می باشد. همانطور که در فصل۳ اشاره کردیم، ورودی های GenBank، EMBL و DDBJ زیادی می توانند به یک توالی پروتئینی خاص مربوط باشند. بر روی لینک CoDingSequence کلیک نمایید، بدین ترتیب جستجو در سرور EBI SRS برای پیدا نمودن CDS[3] ورودی های مربوطه آغاز می شود.
- ✓ **PIR:** این فیلد شامل شماره های دسترسی ورودی های معادل در منبع قدیمی Protein Information Resource (PIR) می باشد، که اکنون به کار خود ادامه نمی دهد، اما با کنسرسیوم UniProtKB ادغام شده است.
- ✓ **UniGene:** لینکی برای اتصال به بانک اطلاعاتی بیان ژن در NCBI (برای دسترسی به اطلاعات آزمایشگاهی در مورد DNA chip ، به فیلد CleanEx رجوع نمایید).
- ✓ **PDB:** این فیلد شامل یک لینک برای اتصال به یک توالی هومولوگ با توالی مورد نظر ما، که اطلاعاتی در مورد ساختار سه بعدی آن در دست است، می باشد. در مورد مثال ما، لینک به ورودی P113 2 در the Swiss-Prot) (basic fibroblast growth facror 1 در این قسمت قرار دارد،که ساختارهای کریستالی متعددی از این

[1] dimerization
[2] National Institute of Health
[3] *CoD*ing Segment، قسمتی از توالی نوکلئوتیدی است که بطور اختصاصی پروتئین را از ابتدا تا انتهاکد می کند.

ورودی در بانک اطلاعاتی پروتئین (PDB) موجود است. بر روی لینک PDB کلیک نمایید تا صفحه مربوطه را مشاهده نمایید (کد شناسایی PDB در اینجا 1FGK می باشد).

- ✓ **ModBase**: این فیلد به منظور اتصال به ModBase می باشد. در این بانک اطلاعاتی بانک اطلاعاتی مدل ها به صورت تئوری محاسبه شده اند، نه بر اساس یافته های تجربی. این مدل ها ممکن است حاوی اشتباهات فاحشی باشند، که اغلب نویسنده ها به آن اقرار می کنند. در فصل ۱۱، کار های بی شماری را که با در اختیار داشتن اطلاعات ساختاری همولوگ های پروتئین مورد نظر می توانید انجام بدهید، به تفصیل شرح داده ایم.

نکته حائز اهمیت این است که در صورت عدم وجود اطلاعات ساختاری (به صورت مستقیم یا غیر مستقیم) فیلد PDB هم وجود نخواهد داشت، اما در این مثال چون با یک ورودی بزرگ Swiss-Prot کار می کنیم، این فیلد را مشاهده می کنید.

- ✓ **DIP and IntAct**: این فیلد، شما را به بانک های اطلاعاتی بر هم کنش های پروتئین– پروتئین متصل می نماید. DIP معرف *D*atabase of *I*nteracting *P*roteins است و معرف بانک اطلاعاتی است که توسط UCLA حمایت می شود. این فیلد از آن جهت منبعی حائز اهمیت است که لیستی از تمامی پروتئین هایی را که بصورت تجربی واکنش آنها به وسیله رسپتور EGF پروتئینی نشان داده شده است، در خود جای داده است. IntAct (پروژه ای از سوی EBI)، بر اساس اطلاعات یافته شده در منابع علمی بنا نهاده شده است.
- ✓ **GlycoSuiteDb**: این فیلد تنها در صورتی وجود خواهد داشت که پروتئین مربوطه یک الگوی گلیکوزیله شدن شناخته شده را داشته باشد. در این فیلد لینکی برای دسترسی به ساختارهای شیمیایی بخش گلیکوزیدی وجود دارد. سرویس رایگان این سایت تنها برای دانشگاه ها فراهم بوده و برای استفاده از آن نیاز به ثبت نام است، که این خود یک مانع بزرگ برای محققین است.
- ✓ **SWISS-2DPAGE**: این فیلد شامل لینکی برای دسترسی به دنیای پروتئومیک می باشد. با کلیک بر روی لینک SWISS-2DPAGE به مجموعه ای از الکتروفورز های ژل دوبعدی(2D –gel) از تومورها و بافت های انسانی مختلف دسترسی پیدا می کنید. اگر پروتئین مورد نظر بر روی ژل تشخیص داده شده باشد، نقاطی به صورت مشخص درآن وجود دارد. در مورد ورودی P00533 این نقطه ها وجود دارند. در غیر اینصورت مجموعه ای از جایگاه های مورد انتظار روی ژل، بر اساس نقطه ایزوالکتریکی به صورت تئوری (PI) و وزن مولکولی پیش بینی شده نشان داده خواهند شد.
- ✓ **HGNC**: این فیلد (تنها برای پروتئین های انسانی) لینکی برای دسترسی به نام های رسمی ژنهای انسانی که توسط HUGO (سازمان ژنوم انسانی، کمیته نام گذاری ژن) فراهم شده است، ارائه می کند..
- ✓ **GeneCards**: این فیلد لینکی را برای دسترسی به ورودی معادل در بانک اطلاعاتی GeneCards که از طرف Weizmann Institute of Science (در اسرائیل) حمایت می شود، به شما معرفی می نماید.
- ✓ **GeneLynx**: این فیلد شامل لینکی برای دسترسی به ورودی معادل در GeneLynx (مجموعه ای از لینک های دسترسی سریع به منابع در دسترس عموم در ارتباط با هر ژن انسانی) می باشد.
- ✓ **MIM**: این فیلد (بخصوص برای پروتئین های انسان و موش) شامل لینکی برای دسترسی به ورودی معادل OMIM (بانک اطلاعاتی بیماری های ژنتیکی انسان با نام Online Mendelian Inheritance in Man) می باشد.

- ✓ **Ontologies**: این بخش لیستی از تمامی عملکرد های محصولات ژن EGFR را، با رعایت سلسله مراتب و رتبه بندی ها، در اختیارتان قرار می دهد. با کلیک بر روی لینک QuickGo view (در قسمت Send از این فیلد)، خلاصه ای زیبا از تمامی اطلاعات بدست آمده نشان داده خواهد شد.
- ✓ **Profiles/domains**: فیلد Cross-References شامل مجموعه ای از فیلد ها می شود که با دسته بندی پروتئین ها بر اساس خانوادهایی از توالی های مرتبط، بر اساس دامین های اصلی آنها، سر و کار دارند.

سیستم های دسته بندی مختلفی بصورت همزمان وجود دارند. این سیستم ها هر کدام از الگوریتم ها یا روش های متنوعی برای توصیف و ردیابی دامین های تکرار شونده در توالی یا ساختار پروتئین و همچنین دسته بندی پروتئین ها به گرو های مختلف، استفاده می کنند. اصول کلی حاکم بر این دسته بندی ها و نیز موارد استفاده آنها در فصول ۷، ۹ و ۱۱ به تفصیل آورده شده است. کلید واژه های مربوطه به شرح زیر می باشند:

- InterPro
- Pfam
- Prodom
- PRINTS
- SMART
- PROSITE
- BLOCKs

اگر زمان لازم برای تنها یک کلیک را دارید، از لینک InterPro استفاده نمایید، این لینک شامل خلاصه ای از لینک های دیگر است (البته به جز BLOCKs) . با کلیک بر روی لینک Graphical view of domain structure، در ردیف InterPro، بلافاصله دامین های متفاوت تکرار شونده بسیاری را که در پروتئین رسپتور EGF با کد P00533 شناسایی شده اند، ملاحظه خواهید کرد.

- ✓ **Ensembl**: این فیلد تنها برای پروتئین های انسان و موش طراحی شده است و شامل یک لینک به پرونده ژنی معادل در بانک اطلاعاتی Ensembl (یک منبع اصلی ژنوم انسانی) می باشد (برای اطلاعات بیشتر در مورد Ensembl به فصل ۳ مراجعه کنید).

اعتقاد ما بر این است که این مجموعه لینک ها در Swiss-Prot، با داشتن لینک های PDB و NCBI/PubMed، از سودمند ترین مجموعه های لینک در زمینه ژن های پستانداران می باشد.

- ✓ **RZPD-ProtExp and SOURCE**: این فیلد های مفید (به خصوص برای آزمایشگر ها) در بر گیرنده منابع عمومی برای تهیه مواد مورد نیاز (مانند کلون های cDNA) برای کار بر روی ژن EGFR و یا پروتئین حاصل از آن است. در صورت کلیک بر روی لینک های مربوطه، حتماً تاکنون به اطلاعات مفیدی در مورد پروتئین خود دست یافته اید.

کلید واژه ها

این بخش از ورودی در بالای شکل ۴-۴ دیده می شود. این فیلد لیستی از اصطلاحات مربوط به پروتئین مورد نظر از قبیل transmembrane، receptor، glycoprotein و نظایر آن را در اختیارتان قرار می دهد. با کلیک بر روی هر یک از آنها لیستی از تمامی ورودی های Swiss-Prot حاوی این کلید واژه بدست می آید. با افزایش روز افزون حجم این بانک

اطلاعاتی، ما مطمئن نیستیم که دیگر این روش جستجو مفید و کاربردی باشد. برای مثال واژه کلیدی receptor در حدود بیش از ۱۵۷۰ ورودی را در اختیار می گذارد!

با استفاده از گزینه های Advanced search و Full text-search در سیستم جستجوی SRS در سرور ExPASy (فصل ۲)، می توانید در Swiss-Prot جستجوی ترکیبی کلید واژه ها را انجام بدهید. به این ترتیب می توانید جستجو های اختصاصی تری انجام دهید.

بخش Features

در قسمت های پایینتر ورودی P00533، به بخش شاخصه ها (Features) وارد می شوید. در این بخش شاخصه های پروتئین، بصورت یک نقشه روی توالی نشان داده شده است(فصل ۳). در صورتی که با Genbank کار کرده باشید، این مفهوم برای شما ملموس تر خواهد بود. البته خوشبختانه این موضوع درمورد پروتئین ها بسیار ساده تر است.

شکل ۴-۴ ابتدای بخش Features را برای ورودی P00533 نشان می دهد.

این بخش به شکل جدولی با سرعنوان های زیر سازماندهی شده است:

Key, From, To, Length, Description

فیلد Key مشخص کننده نوع شاخصه ای است که اکنون پیش روی شماست. اعداد ذکر شده در فیلد های From و To مربوط به شماره گذاری توالی (از انتهای N به انتهای C) می باشد، در حالیکه فیلد Length، طول توالی را بیان می کند. در قسمت Description خلاصه ای مفید از شاخصه مورد نظر آورده شده است.

با برگشت به ستون فیلد Key مربوط به ورودی P00533 (شکل ۴-۴) می توان اعداد مربوط به اصطلاحات کلیدی زیر را مشاهده نمود:

- ✓ **SIGNAL:** اعداد مرتبط با این اصطلاح کلیدی، مشخص می کنند اسید آمینه های ۱ تا ۲۴ یک پپتید سیگنالی هستند، که ویژگی یک رسپتور پروتئینی عبور کننده از عرض غشاء است.

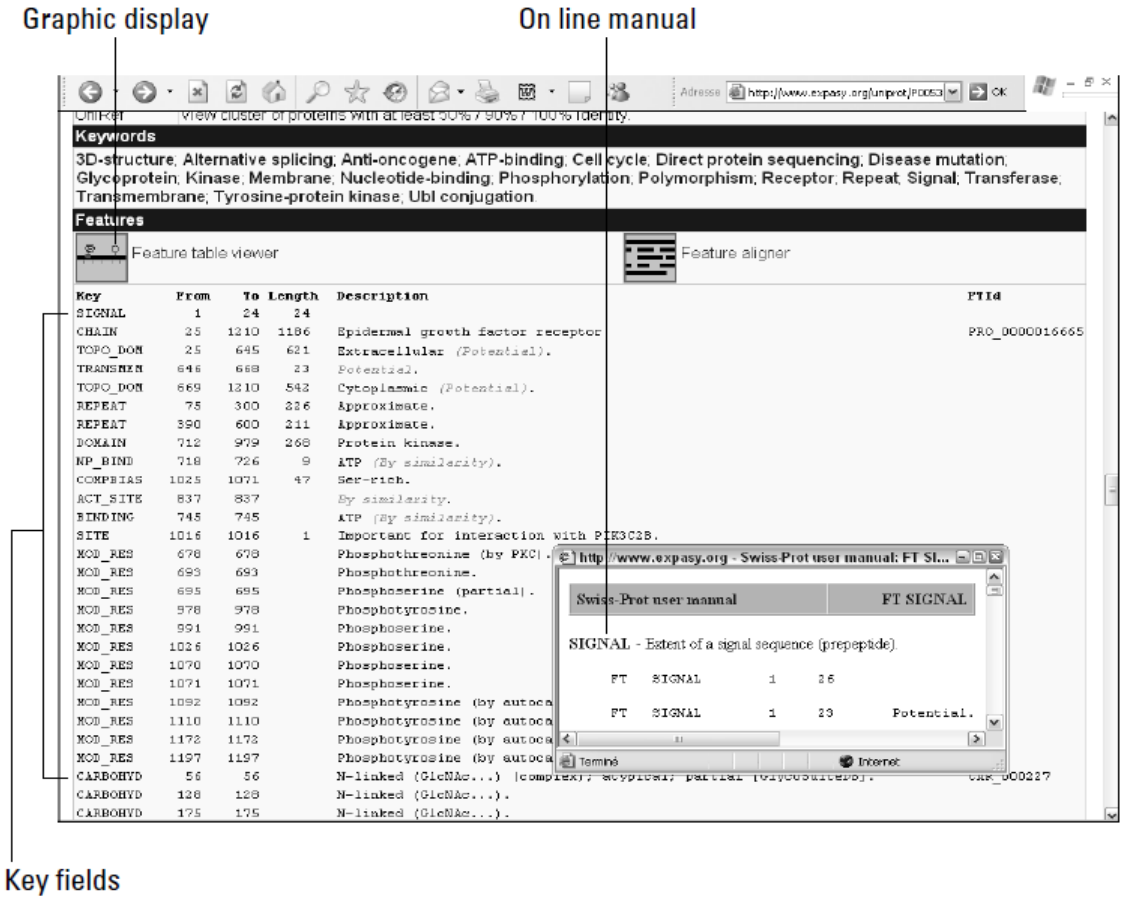
توجه کنیدکه با کلیک برروی لینک SIGNAL (مشابه سایر لینک های مربوط به کلید واژه ها در این فیلد)، بخش مربوطه در راهنمای استفاده Swiss-Prot (شکل ۴-۴) را به شما ارائه می دهد. با خواندن این مطالب می توانید به درک صحیحی از معنی این اصطلاح دست یابید. با کلیک بر روی رنج مورد نظر از توالی که زیر آن خط کشیده شده است (برای مثال 1 24)، این قطعه در توالی بصورت مشخص نشان داده می شود.

- ✓ **CHAIN:** در این قسمت می توان زنجیره پپتیدی بالغ را از اسید های آمینه ۲۵ تا ۱۲۱۰ (از انتهای c) مشاهده نمود، که نشاندهنده اینست که در حقیقت پپتید سیگنالی از پروتئین بالغ جدا شده است.
- ✓ **TOPO_DOM**: معرف Topological domain است. این قسمت نشاندهنده اینست که کدام قسمت از توالی در خارج سلول[1] و کدام قسمت در داخل سلول[2] قرار خواهد گرفت.
- ✓ **TRANSMEM**: در این قسمت مشاهده می شود که ۲۳ اسید آمینه بعدی (8 – 4) مربوط به بخش عبور کننده از غشای پروتئین هستند.

[1] Extracallular domain
[2] Intracellular domain

✓ **Domain**: این اعداد نشاندهنده یک دامین دیگر هستند. اسید آمینه های ۶۶۹ تا ۱۲۱۰ که به طور کامل در داخل سلول قرار دارند، دامین داخل سلولی پروتئین را تشکیل می دهند.

توجه نمایید که آنچه Swiss-Prot به آن Domain می گوید در حقیقت موقعیت وسیع تری را نسبت به آنچه که منظور متخصصان معمولاً در بانک های اطلاعاتی چون InterPro و یا Pfam است، را در بر می گیرد. برای نمونه می توان به دامین خارج سلولی پروتئین EGFR اشاره نمود که در InterPro شامل چندین دامین می شود (برای مشاهده بر روی لینکها کلیک نمایید).



شکل ۴-۴: بخش فوقانی Features مربوط به ورودی Swiss-Prot P00533

کلید واژه Domain برای اشاره به قطعه غنی از سرین و ناحیه ای از توالی که مسئول فعالیت پروتئین کیناز است، نیز بکار می رود. در حقیقت این امر تأکیدی است بر آنچه پیش از این پیرامون استفاده گسترده از این عنوان در Swiss-Prot گفته شد. بیایید جستجوی خود را با اصطلاحات جدیدتر ادامه دهیم:

✓ **NP-BIND**: این اعداد در حقیقت گستره "ناحیه اتصال فسفونوکلئوتیدها" (اسید آمینه های ۷۱۸ تا ۷۲۶) برای اتصال به ATP می باشند.

✓ **BINDING**: در این مکان، ناحیه دقیق اتصال ATP، بر روی لیزین در موقعیت ۷۴۵، نشان داده شده است.

✓ **ACT_SITE**: اعداد در این مکان آمینواسید های مرتبط با فعالیت یک آنزیم را نشان می دهند.

برای افراد غیر حرفه ای سه مکان ذکر شده در بالا دارای مفهوم تقریباً مشابهی هستند.

- ✓ **COMPBIAS**: گستره ناحیه ای را که از نظر نوع اسید های آمینه سو گیری[1] دارد (در این مورد ناحیه غنی از سرین) را نشان می دهد.

عناوین بعدی حاوی اطلاعاتی پیرامون اسیدهای آمینه است که بعد از ترجمه دچار تغییرات شیمیایی شده اند (شکل ۵-۴):

- ✓ **MOD_RES**: این اعداد نشاندهنده مناطقی است که فسفریلاسیون[2] بر روی آنها صورت گرفته است. بر روی لینک **MOD_RES** کلیک نمایید، لیستی از تغییرات شیمیایی مرتبط با کلید واژه مورد نظر در اختیارتان قرار خواهد گرفت.
- ✓ **CARBOHYD**: در این مکان، تعداد زیادی اسید آمینه را که بر روی آنها مولکول های کربوهیدرات مستقر شده اند، و همچنین نحوه اتصال آن ها (–C: اتصال از طریق کربن، –N: اتصال از طریق نیتروژن و –O: اتصال از طریق اکسیژن) ، نشان داده شده است. دقت داشته باشید که این ها همگی در دامین خارج سلولی قرار دارند. با کلیک بر روی کلید واژه CARBOHYD ، جزئیات و پیچیدگی های بیشتری در موردگلیکوزیلاسیون به شما نشان داده خواهد شد.
- ✓ **DISULPHID**: این اعداد نشاندهنده جفت سیستئین های اتصال یافته در پروتئین نهایی را نشان می دهد. اینجا تعداد زیادی از این اتصالات را می توانید ببینید.

در نهایت، بخش Features با لیستی از مشخصات تخصصی پایان می پذیرد:

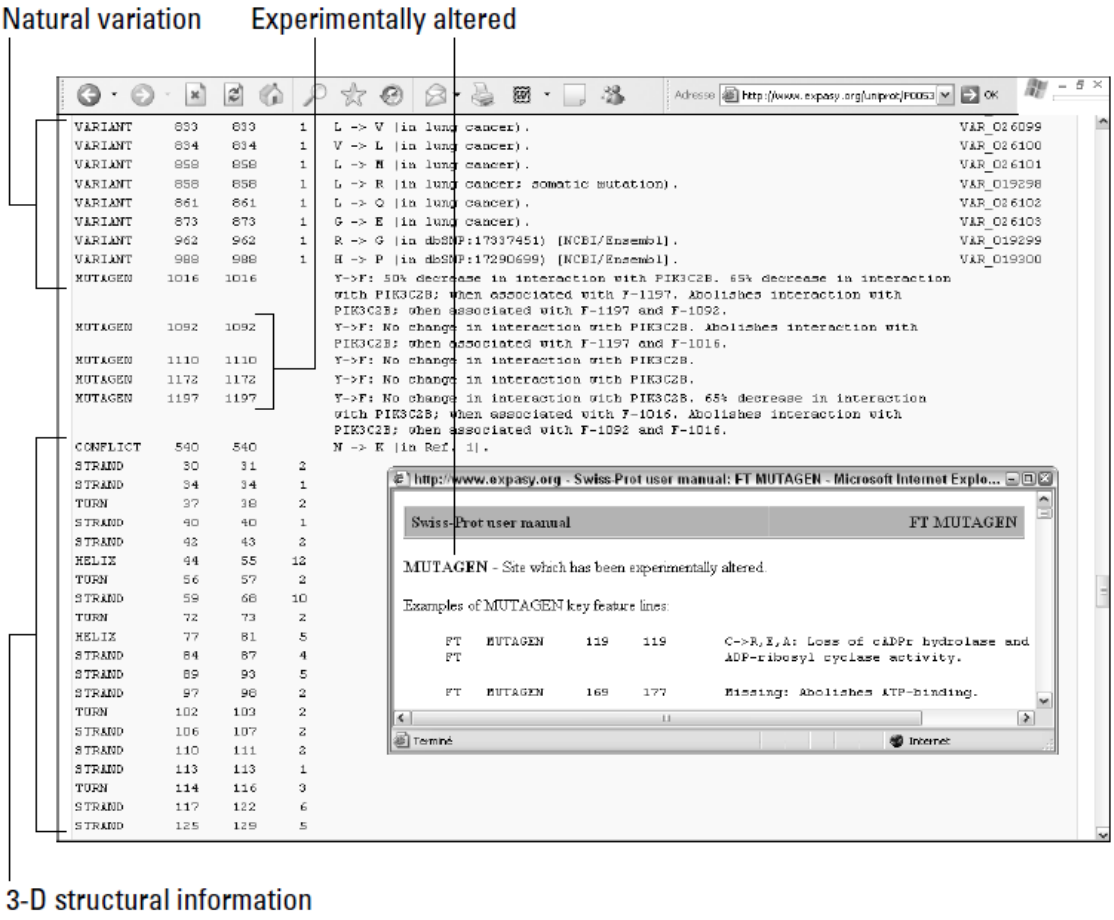
- ✓ **VAR_SEQ**: این فیلد نشاندهنده تغییرات آمینواسیدی است که از ترجمه mRNA های متفاوت (ایزوفرم ها) ناشی از طریقه های متفاوت اتصال، ایجاد می شوند.
- ✓ **VARIANT**: این مکان نشاندهنده گونه های طبیعی توالی پروتئین EGFR می باشد، که بیشترآن ها با سرطان ریه مرتبط می باشند.
- ✓ **MUTAGEN**: بر خلاف فیلد قبل، این فیلد تغییرات توالی که در آزمایشگاه و به صورت عمدی بر روی پروتئین ها صورت گرفته است، را ثبت می کند.
- ✓ **CONFLICT**: این مشخصه، به منظور نمایش تفاوت های موجود بین منابع گوناگون یک توالی پروتئین خاص، بخصوص خطاها و یا پلی مورفیسم های ناشناخته، می باشد.

در ادامه بخش شاخصه ها به بخش توصیف اشکال و ساختارهای ثانویه پروتئین وارد می شویم. ساختارهای ثانویه در سه حالت نمایان می شوند: STRAND (ساختار نخ مانند و گسترده) ، TURN (از نامش مشخص است) و HELIX (ساختار پیچ خورده و فنر مانند).

قبل از اتمام بخش شاخصه ها به آیکن Feature Table Viewer در سمت چپ صفحه نمایش، نگاهی بیاندازید (شکل ۴-۴، سمت چپ و بالا). بر روی آیکن کلیک نمایید، صفحه ای جدید باز می شود که یک تصویر گرافیکی مشابه شکل ۶-۴ را نشان می دهد. این تصویر گرافیکی، ساختار اولیه پروتئین EGFR را نشان می دهد. اگر با موس روی شکل بروید، مشخصات مرتبط با هر سمبل را به شما نشان می دهد. در مورد P00533، این کار به شما نشان می دهد که تمامی تغییرات فسفریلاسیون در دامین داخل سلولی اتفاق می افتد، حال آنکه دیگر تغییرات (اتصالات دی سولفیدی و گلیکوزیله شدن) در دامین خارج سلولی صورت می گیرد. در صورتیکه اندکی در زمینه بیوشیمی کار کرده باشید، خواهید فهمید که این مطلب کاملا قابل درک بوده و منطقی به نظر می رسد.

---

[1] bios
[2] Phosphorilation



شکل ۵-۴: بخش انتهایی شاخصه ها در ورودی P00533

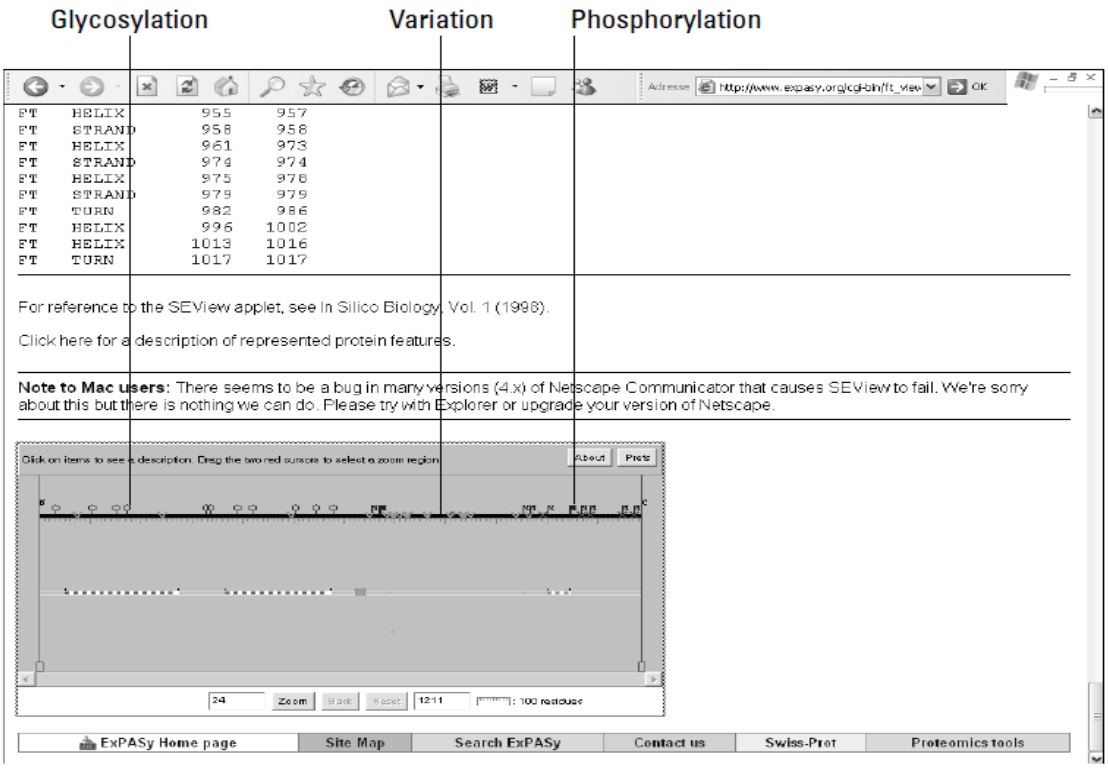
در نهایت، خود توالی

مطلب خاص قابل ذکر در مورد این بخش وجود ندارد. به خاطر داشته باشید که می توانید با کلیک بر روی P00533 in FASTA format که در سمت راست و پایین ورودی مربوطه قرار دارد، توالی را به فرمت FASTA تبدیل کنید. می توانید توالی را از طریق گزینه Save As در منوی File و یا با Cut و Paste کردن آن در یک محیط Word، ذخیره نمایید.

اطلاعات بیشتر پیرامون پروتئین انتخابی شما

پس از مطالعه دقیق ورودی Swiss-Prot پروتئین مورد نظر خود (و یا همولوگ آن) و نیز استفاده از تمامی لینک های موجود در آن، شما احتمالاً با ۹۰٪ از آنچه که همه دنیا در مورد آن پروتئین می دانند، آشنا می شوید. قدم بعدی، درک و فهم منسجم و یکپارچه کردن مطالبی است که در Swiss-Prot و یا در لینک های آن خوانده اید. برای مثال، احتمالاً مفهوم دقیق myristilation را هنوز درک نکرده اید و یا در حد کافی در مورد واکنش های بیوشیمیایی کاتالیز شده توسط پروتئین نمی دانید و یا می خواهید بدانید یک آنزیم در چه مسیر متابولیسمی دخیل است.

بطور خلاصه، اکنون زمان آن رسیده است که به اطلاعات اضافی، برای فهم بیشتر آنچه در Swiss-Prot با آن مواجه شده اید، نیاز پیدا کنید. در ادامه مبحث تعدادی از منابع بسیار ارزشمند برای خوانندگان علاقمند ارائه خواهد شد.



شکل ۶-۴: بخش Feature table viewer مربوط به ورودی P00533 در Swiss-Prot

یادگیری بیشتر پیرامون آمینواسیدهای تغییر یافته

اگر می خواهید در رابطه با تغییرات اعمال شده بر روی آمینواسیدها در هنگام بالغ شدن پروتئین بیشتر بدانید، به بانک اطلاعاتی RESID[(r)] نگاهی بیاندازید. RESID در حقیقت بانک اطلاعاتی تغییرات پس از ترجمه می باشد که توسط John Garavelli در مؤسسه بیوانفورماتیک اروپا[1] مدیریت می شود. برای مثال، شما می توانید از این منبع برای درک دقیق مفهوم myristylation استفاده نمایید. برای این منظور دستورالعمل زیر را دنبال کنید:

۱. به آدرس اینترنتی www.ebi.ac.uk/RESID/ مراجعه کنید.
بانک اطلاعاتی تغییرات پروتئین RESID نمایان خواهد شد.
۲. بر روی لینک Search RESID کلیک نمایید.
فرم جستجو ظاهر خواهد شد.
۳. در پنجره جستجو myristoylation را وارد نمایید (شکل ۷-۴).
۴. بر روی دکمه Search کلیک نمایید.

نتیجه جستجو جدولی حاوی سه ورودی خواهد بود: AA0059 (N-myristoyl-glycine)، AA0078 (N - myristoyl-L-lysine), و AA0307 (S-myristoyl-L-cysteine)

۵. در ستون انتهای سمت چپ، با عنوان RESID ID،بر روی لینک AA0078 کلیک نمایید.

اطلاعات کاملی پیرامون این آمینواسید تغییر یافته در اختیارتان قرار خواهد گرفت، که بخش پایانی آن فرمول شمیایی آمینو اسید است.

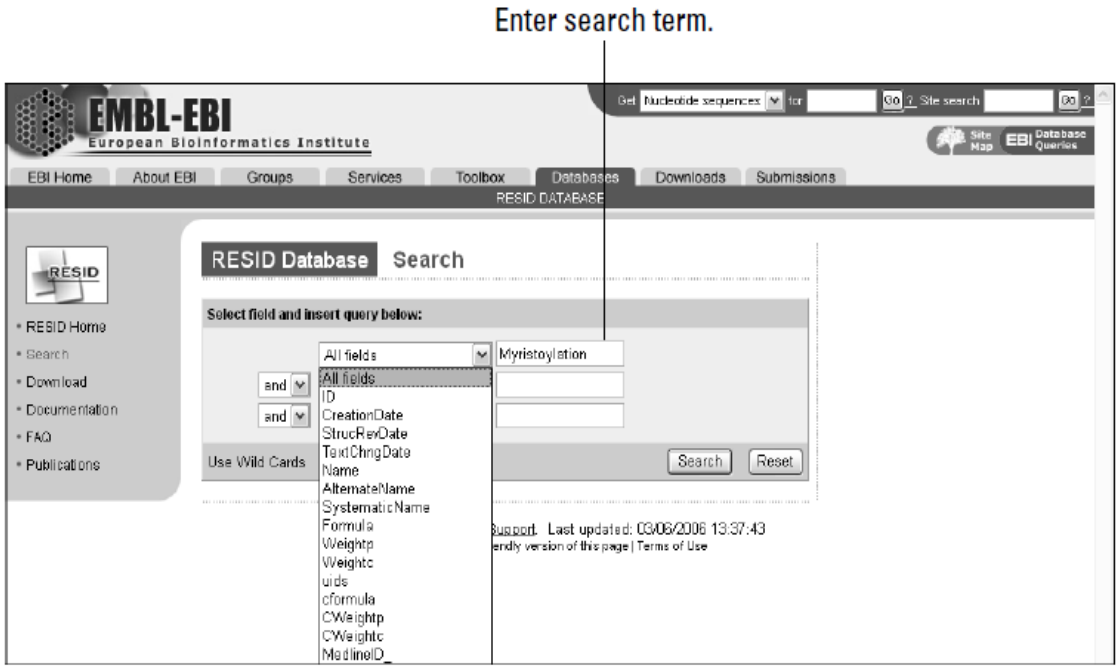
[1] European Bioinformatics Institute

معرفی چند سایت تخصصی و پیشرفته بیوشیمی

اگر قصد دارید مهارت های شیمی و بیوشیمی خود را ارتقاء بدهید، به سایت های زیر سری بزنید:

- ✓ **The Glycan Structure Database:** به آدرس www.glycosuite.com/ .

این سایت تنها برای دانشگاهها رایگان است و نیاز به ثبت نام دارد.



شکل ۷-۴: جستجو در بانک اطلاعاتی RESID برای یافتن Myristoylation

- ✓ **The Lipid Bank:** به آدرس lipidbank.jp
- ✓ **ChemIDplus:** به آدرس chem.sis.nlm.nih.gov/chemidplus/.

این بانک اطلاعاتی خیلی جالب است، در این بخش می توانید از طریق رسم مولکول خیالی خود! در بانک اطلاعاتی جستجونمایید.

توجه کنید که استفاده از این قبیل بانک اطلاعاتی برای غیر متخصصین گمراه کننده است. هدف ما در این قسمت این است که شما بدانید چنین منابعی هم وجود دارند. قبل از اینکه بتوانید مسیر های بیوشیمیایی را در این منابع جستجو کنید، باید کمی با آن ها آشنایی داشته باشید.

اطلاعات بیشتر پیرامون مسیرهای بیوشیمیایی

آگاهی از اینکه پروتئین انتخابی شما در مسیر بیوسنتز di-amino-pymelate شرکت دارد، شاید مفهومی برای شما نداشته باشد! برای اینکه بدانید پروتئین شما در چه مسیر بیوشیمیایی شرکت دارد، به سایت Boehringer سری بزنید. مسیر های بیوشیمیایی، یک توالی از واکنش های اولیه هستند که منجر به تولید ترکیبات ضروری برای سلول می شوند. اگر تاکنون

مطالعه ای در زمینه بیوشیمی انجام داده باشید، حتماً چارت مسیر های بیوشیمیایی را بر روی دیوار گروه بیوشیمی مشاهده نموده اید. آیا می دانید که امروزه می توانید بطور اتوماتیک این مسیر ها را جستجو نمایید؟

۱. به آدرس اینترنتی www.expasy.ch/cgi-bin/search biochem-index مراجعه نمایید.
۲. کلمه Pimelate را در پنجره Keyword Search تایپ نمایید و سپس بر روی گزینه Submit Query کلیک نمایید.

صفحه Pimelate ظاهر می گردد. pimelate که در واقع مولکول ضروری برای ساخت دیواره سلولی باکتری می باشد، تنها به عنوان مثال آورده شده است. در صورت تمایل می توانید آنرا با نام ترکیبی که بر روی آن کار می کنید جایگزین نمایید.

۳. در زیر DIAMINOPIMELATE- (اولین ورودی در لیست)، بر روی J3 کلیک نمایید.

در یک چارت ظاهر می شود که نشان می دهد که این ترکیب برای بیوسنتز دیواره سلولی باکتریایی ضوررری است.

جدول ۱-۴ لیستی از چند منبع دیگر در مورد مسیر های بیوشیمیایی ارائه می کند.

توضیحات	آدرس اینترنتی
دائره المعارف مشهور ژنها و ژنوم ها (KEGG[1])، کدهای الکترونیکی E.C. و یا نام های ژن بهترین گزینه برای شروع کار با این منبع می باشد.	www.genome.ad.jp/kegg/
BRENDA، سیستم اطلاعات جامع آنزیم ها[2]، در صورتیکه تمایل به تجربه آزمایشات واقعی را دارید سری به آن بزنید.	Brenda.bc.uni-koeln.de
سایت رسمی مجموعه نامگذاری آنزیم ها[3] مرتبط با پایگاه بین المللی بیوشیمی و بیولوژی مولکولی (IUBMB[4]).	www.chem.qmul.ac.uk/iubmb/
دائره المعارف ژنها و متابولیسم Ecoli که در حال اضافه کردن اطلاعات سایر باکتری ها نیز می باشد.	www.ecocyc.org/

جدول ۱-۴: منابع مسیر های بیوشیمیایی آنلاین

یافتن اطلاعات بیشتر در مورد ساختارهای پروتئینی

اگر توالی پروتئین انتخابی خود را از طریق تمامی امکانات در دسترس آنالیز نموده اید (جستجوی موتیف ها در فصل ۶، جستجو ی بانک های اطلاعاتی در فصل ۷ و تطابق چندگانه توالی ها در فصل ۹)، اکنون احتمالاً می خواهید بدانید مکان دقیق چند آمینواسید خاص در ساختار سه بعدی پروتئین مربوطه کجاست؟ آیا آن آمینو اسید ها در قسمت های داخلی مولکول قرار دارد و یا بر روی سطح است؟موقعیت آن آمینواسیدها نسبت به آمینواسیدهای دیگر چگونه است؟ آیا در داخل جایگاه فعال آنزیم قرار دارند؟ و پرسش های دیگر. تمامی این پرسش ها نیاز به اطلاعات ساختاری از پروتئین مورد نظر دارد.

[1] Kyoto Encyclopedia of Genes and Genomes (KEGG)
[2] The Comprehensive Enzyme Information System BRENDA
[3] Enzyme Nomenclature
[4] International Union of Biochemistry and Molecular Biology

در فصل ۱۱ یاد می گیرید که چگونه بین توالی ها و اطلاعات ساختاری ارتباط برقرار کنید. اما اکنون جدول ۲-۴ لیستی از منابع مهم را در رابطه با این موضوع در اختیارتان می گذارد.

توضیحات	آدرس اینترنتی
PDB، بانک اطلاعاتی رسمی ساختار های سه بعدی پروتئین	www.rcsb.org/pdb
حاوی ابزار هایی برای نمایش ساختار ها و آنالیز های مقایسه ای	www.ncbi.nlm.nih.gov/structure
Scop، دسته بندی پروتئین ها بر اساس ساختار، که بیان توضیحاتی پیرامون ارتباطات ساختاری و تکاملی میان پروتئین هایی که ساختارشان مشخص شده است، بیان می کند.	Scop.mrc-lmb.cam.ac.uk/scop
CATH (خانواده ای متشکل از *C*lass، Architecture، *T*opology و *H*omologous) دسته بندی رتبه ای از ساختار های پروتئین-دامین را در اختیارتان می گذارد.	www.cathdb.info
Swiss-Model، یک سرور تمام اتوماتیک بوده که همسانی ساختارهای پروتئینی را مدل سازی می نماید و از طریق سرور ExPASy قابل دسترسی است.	Swissmodel.expasy.org

جدول ۲-۴: بانک های اطلاعاتی اختصاصی برای اطلاعات ساختاری

یافتن اطلاعات بیشتر پیرامون خانوده های اصلی پروتئین ها

با وجود گستردگی بسیار زیاد خانواده های پروتئینی، بانک های اطلاعاتی عمومی پروتئین ها، اغلب در فراهم سازی اطلاعات و جزئیات کامل برای کمیته های تخصصی و تحقیقاتی ناتوان می باشند. اما منابع اطلاعاتی بسیار تخصصی برای پر نمودن خلأ های موجود وجود دارند. اگر پروتئین شما به یکی از دسته بندی هایی که در ادامه ذکر می شوند، تعلق دارد، ممکن است نیاز داشته باشید به منابع ارائه شده در جدول ۳-۴ نگاهی بیاندازید.

توضیحات	آدرس اینترنتی
IMGT، بانک اطلاعاتی بین المللی ایمنوژنتیک[1] که در خصوص ایمنوگلوبولین ها، گیرنده های سلول های T و	Imgt.cines.fr

[1] International Immunogenetics database

مولکول های پیچیده که در سازش بین بافت ها نقش دارند[1]، در مهره داران اطلاعاتی را ارائه می دهد.	
REBASE، بانک اطلاعاتی آنزیم های محدود کننده و تغییر دهنده[2] است.	Rebase.neb.com
CAZy، منبع اطلاع رسانی آنزیم هایی است که باعث تغییر، فروپاشی و یا ساخت پیوند های گلیکوزیدی می شوند.	www.cazy.org/CAZY/
MEROPS، بانک اطلاعاتی تأمین کننده اطلاعات بسیار غنی پیرامون پروتئاز ها می باشد.	Merops.sanger.ac.uk
PKR (منبع اطلاع رسانی پروتئین کیناز[3])، مختصری اطلاعات قابل دسترس از طریق وب، پیرامون خانواده آنزیمی پروتئین کیناز ارائه می دهد.	www.kinasenet.org/pkr/
Nursa، اطلس سیگنال های رسپتور هسته[4]، یک منبع اطلاعاتی در مورد رسپتور های هسته ای (برای مثال: رسپتورهای استروئیدی) به همراه فعال کننده ها، مهار کننده ها و لیگاندهای آن ها می باشد.	www.nursa.org
بانک اطلاعاتی مغز انسان[5]، که اطلاعاتی پیرامون پروتئین های شرکت کننده در فرآیند های عصبی، مانند کانال های یونی، گیرنده های غشایی انتقال دهنده های عصبی و نیز تعدیل کننده های عصبی[6] و رسپتورهای بویایی، ارائه می دهد.	Senselab.med.yale.edu/senselab
بانک اطلاعاتی COG (خوشه های گروه های ارتولوگ[7])، دسته بندی دوباره پروتئین های دارای حداقل سه دودمان اصلی مشترک، که در واقع دارای دامین های محافظت شده مشترک هستند.	www.ncbi.nlm.nih.gov/COG

جدول ۳-۴: بعضی از بانک اطلاعاتی تخصصی پروتئین

[1] histocompatibility complex molecules
[2] restriction/modification enzymes
[3] Protein Kinase Resource
[4] Nuclear Receptor Signaling Atlas
[5] Human Brain Database
[6] neuromodulators
[7] Clusters of Orthologous Groups

فصل پنجم

کار با توالی منفرد DNA

در این فصل می خوانیم:

- ✓ شناسایی مشکلات کار با توالی مورد نظر
- ✓ پردازش و نمایش یک نقشه محدود (برای نمایش رابطه بین دو پدیده در سطح مولکولی)
- ✓ تعیین خصوصیات توالی مورد نظر
- ✓ یافتن ORF ها، اگزون ها و ژنها
- ✓ گردآوری قطعات یک توالی

.....

" نه هرآنچه قابل شمارش است، شمرده می شود و نه هرآنچه شمرده می شود، قابل شمردن است."
— Albert Einstein (1879–1955)

پروتئین ها اغلب عملکرد های بیولوژیکی را انجام می دهند، لذا زیست شناسان به توالی های DNA توجه کمتری می کنند. در حقیقت ۹۰ درصد DNA مورد استفاده چیزی جز یک وسیله ذخیره کننده اطلاعات نمی باشد، در واقع این ها معادل های زیستی فلاپی و DVD هستند. مولکول DNA شامل اطلاعات بسیاری است که سلول می داند که چگونه از این اطلاعات در جای مناسب استفاده کند، درست مانند یک کامپیوتر که می تواند به یک DVD معنا ببخشد.

در مورد یک پروتئین شما نمی توانید توالی آمینو اسیدی آن را مستقیماً به دست آورید؛ توالی یابی ژن های آن و سپس استنباط کد های ژنتیکی از روی آن بسیار ساده تر است. دقیقاً به همین علت است که توالی های DNA یک مرحله بینابینی و واجب است. شما باید بدانید چگونه از توالی های نوکلئوتیدی برای کار با پروتئین و یا توالی های RNA استفاده کنید. در حقیقت توالی های DNA مادر تمامی توالی ها می باشد!

از اینکه شما در حال مطالعه این کتاب هستید می توان متوجه شد که شما یک دانشمند کارآزموده که در یکی از مراکز گسترده وبی انتهای توالی یابی ژنوم کار می کند و قادر به استنباط نوکلئوتید های خاص از میان میلیونها نوع از آن باشد،

نیستید. این بخش روی آنالیز توالی های نسبتاً کوچک(حد اکثر به طول ۱۰۰ هزار نوکلئوتید) تمرکز دارد. هدف این فصل آموختن مهارت های ضروری در زمینه کار با قطعات ژنومی با اندازه معقول یا با cDNA (DNA کپی شده از روی mRNA) به شما است. cDNA نوعی توالی است که می تواند به سادگی در یک آزمایشگاه بیولوژی مولکولی مجهز ، و یا از طریق یک شرکت توالی یابی و با صرف هزینه ای منطقی تعیین و شناسایی شود.

اولین قدم در کار با یک توالی آنهم به صورت تجربی، اطمینان از صحت آن است، تا یک ماه از وقت خود را بیهوده برای بررسی چیزی که کاملاً غیر قابل استفاده بوده است، تلف نکنید. اولین قسمت از این فصل طریقه پی بردن به مشکلاتی را که در هنگام کار با توالی های جدید با آنها برخورد خواهید کرد، نشان می دهد. همواره قانون Murphy را بیاد داشته باشید که می گوید: اگر امکان ایجاد مشکلی وجود داشته باشد اما تاکنون بحثی در مورد آن نشده است، ممکن است این مشکل برای شما اتفاق بیافتد!

جلوی اشتباه را قبل از آنکه دیر شود بگیرید!

توالی یابی یک قطعه DNA شامل خالص سازی آن، کلون کردن[1] آن توسط یک حامل[2] (مانند پلازمید[3])، تکثیر آن در یک میزبان زیستی (در اغلب موارد یک باکتری مانند *E. coli*) و در نهایت ارائه آن به یکی از پروتکل های توالی یابی مانند روش گسترش پرایمر[4] و یا dye-termination . در حین این پروسه، قطعات DNA ضمیمه عمداً به DNA هدف متصل می باشد. از طرف دیگر، اتفاقات ناخواسته بسیاری ممکن است رخ دهد، که موجب می شود تا توالی مورد نظر آنطور که باید با اطلاعات ژنتیکی مورد مطالعه شما تطابق نداشته باشد. شایع ترین مشکلی که با آن مواجهید این است که توالی مورد نظر شما با توالی هایی از حامل مورد استفاده توسط شما و یا شخص دیگر در آزمایشگاه آلوده شود. شناسایی این مشکل و دانستن نحوه برخورد با آن به اندازه ای مهم است که ما بخش بعدی را دقیقاً به همین موضوع اختصاص داده ایم.

جداسازی و حذف توالی های حامل

مولکول DNA و یا cDNA ای که به منظور توالی یابی فرستاده اید، به یک حامل کلون کننده (پلازمید، فاژ[5]، کازمید[6] و یا YAC ، BAC, PAC) وارد شده و از این طریق تکثیر می یابد. توالی هایی که از این طریق به دست می آیند، ناچاراً دارای قطعاتی از حامل نیز می باشد. شما می توانید آن ها را شناسایی و سپس جداسازی نمایید. این کار به سادگی از طریق جستجوی توالی حامل مورد استفاده توسط توالی های مکمل امکان پذیر است (به عنوان یک محقق مسئول باید این اطلاعات را بدانید).

اما ممکن است توالی های مورد نظرتان با حامل هایی که دیگران استفاده می کنند نیز آلوده شود. در این حالت عاقلانه ترین کار این است نه تنها حامل خود، بلکه حامل های دیگری که احتمال می دهید نمونه شما را آلوده کرده اند را جستجو کنید. برای این کار می توانید از ابزارهایی که در سایت NCBI (مرکز ملی اطلاعات بیوتکنولوژی) موجود است، استفاده نمایید:

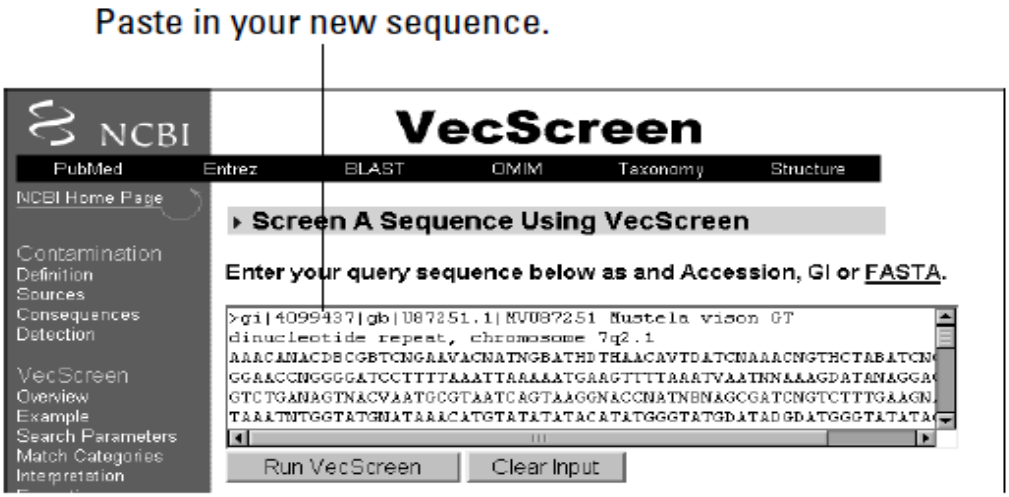
---

[1] cloning
[2] vector
[3] plasmid
[4] primer extention
[5] phage
clasmid

۱. به آدرس اینترنتی www.ncbi.nlm.nih.gov/VecScreen/VecScreen_docs.html مراجعه کنید.
صفحه ورودی VecScreen ظاهر می شود. این صفحه حاوی خودآموزی برای ناخالصی های توالی (برای مطالعه آن روی لینک Contamination کلیک نمایید) و نیز توضیحی از نحوه کار VecScreen می باشد. در حقیقت، این ورودی توالی شما را در یک بانک اطلاعاتی حاوی تمامی حامل ها و توالی های تعدیل کننده کلونینگ شناخته شده به روش blastn (به فصل ۷ مراجعه نمایید) جستجو می کند. این بانک اطلاعاتی UniVec نام دارد. با کلیک بر روی لینک UniVec، صفحه ای با اطلاعات جامع پیرامون این پایگاه بی نظیر در اختیارتان قرار خواهد گرفت. هرگاه احساس آمادگی برای کسب تجربیات واقعی را پیدا نمودید، توالی خود را بردارید و به سراغ مرحله ۲ بروید.

۲. وارد آدرس اینترنتی www.ncbi.nlm.nih.gov/VecScreen/ شوید.
یک پنجره برای وارد کردن توالی باز می شود.

۳. توالی خود را در پنجره مخصوص باز شده وارد نمایید (شکل ۱-۵) .
می توانید از قالب FASTA و یا قالب خام استفاده کنید.



شکل ۱-۵: VecScreen ، پنجره وارد نمودن توالی

۴- بر روی گزینه Run VecScreen در زیر پنجره کلیک نمایید.
صفحه ای با ساختار BLAST نمایان می گردد که نشان می دهد که جستجو را BLAST انجام داده است، اما با نقاب VecScreen.(برای اطلاعات بیشتر در مورد BLAST در فصل ۷ مراجعه کنید).

بهتر از همان ابتدا بجای blastn از VecScreen استفاده نمائید چون VecScreen پارامترهایی را انتخاب می کند که برای شناسائی حامل بهتر باشد.

۵- برروی دکمه Format کلیک نموده و منتظر بمانید.
پارامترهای ابتدائی را تغییر ندهید. بسته به میزان بار سرور BLAST، ممکن است یک صفحه انتظار ظاهر شود، یا بلافاصله نتایج جستجو ظاهر شود.

۵- نتایج جستجو به دو حالت می تواند باشد:

- تشابه معنی داری یافت نشد: این خبر خوبی است. در حقیقت این پیام نشان می دهد که توالی انتخابی به هیچ حامل یا تعدیل کننده شناخته شده ای شبیه نیست. در این صورت شما می توانیدبر اساس ادامه مطالب این فصل، مراحل بعدی آنالیز را انجام دهید.
- موارد مشابهی از جستجویتان یافت شده است: این خبر بدی است. معنی آن این است که توالی شما حاوی چند حامل یا تعدیل کننده بوده است.

گزارش ناخالصی در VecScreen شامل دو بخش است: بخش بالائی نشان دهنده ی چگونگی پراکندگی توالی های ناخالص در طول توالی شما می باشد. این نمایش بصورت تصاوی با کد های رنگی می باشد (شکل ۲-۵، به صورت آنلاین به آن مراجعه کنید تا حالت رنگی آن را ببینید). بخش دوم، نتیجه استاندارد بدست آمده از Blastn می باشد که تعیین کننده ی هویت حامل یا تعدیل کننده متصل شده به توالی شما و نیز alignment دقیق آن است.

رنگ ها در تصاویر بانک اطلاعاتی UniVec نشاندهنده حالت های اتصال شدید (قرمز)، متوسط (بنفش) و یا ضعیف (سبز) هستند.کد زرد به منظور ترسیم قطعاتی است که منشا آن ها مشخص نیست، مثل قطعات کوچک در بین دو حامل متصل شده. US 725، دو منطقه که نشاندهنده تشابه ۱۰۰ درصد با حامل کلونیگ معروف PBR322 است، نمایش می دهد. این مورد یک توالی ناخالص می باشد که قبل از آنکه سیستم فیلتر کننده حامل بصورت سیستماتیک در مراحل پذیرش وارد شود، در GenBank ثبت شده است.

آنچه در مراحل بعد باید بر روی توالی ناخالص انجام دهید در یک متن خودآموز کامل ارائه شده است، که با کلیک بر روی لینک Interpretation of VecScreen Results (در بالای هر نتیجه جستجوی مثبت) قابل دسترسی است.

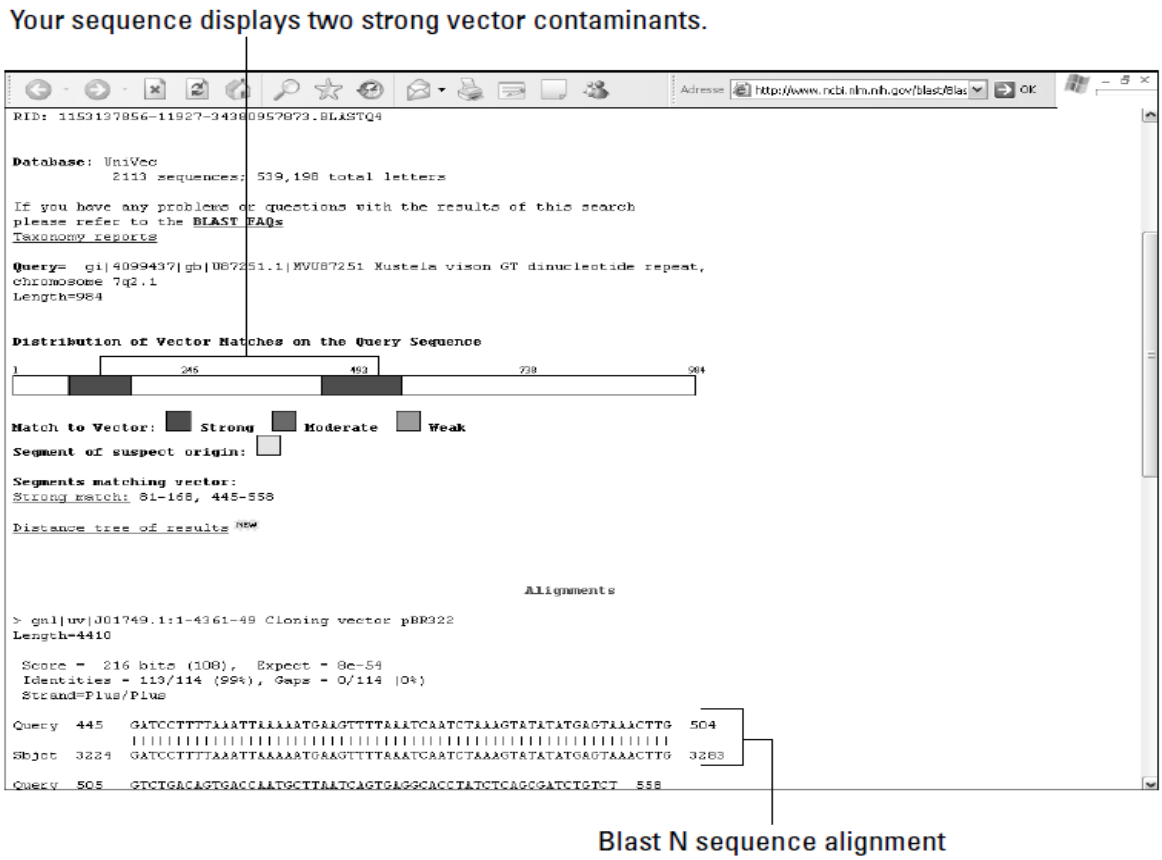
بطور کلی، به هنگام یافتن ناخالصی ها شما دو گزینه پیش رو دارید. اگر ناخالصی ها در انتهای توالی ها وجود دارند و ناشی از حاملی است که استفاده می کنید، کافی است آن بخش ها را حذف کرده و عملیات خود را ادامه دهید. ولی اگر قطعات ناخالص در بخش های دیگر از توالی تان یافت شده است (مطابق شکل ۲-۵) و یا ناخالصی ناشی از حامل خودتان نیست، بهتر است آن توالی را دور بیندازید.

مواردی که نباید توالی تان را دور بیندازید

در موارد مشاهده انطباق در VecScreen، به محض دیدن نام ورودی بانک اطلاعاتی UniVec نتیجه نگیرید که نمونه شما با وکتورهای دیگران آلوده بوده است. شما ممکن است از pUC 19 به عنوان حامل کلونیگ استفاده نموده و آنچه را که VecScreen از یک انطباق قوی با pUR 222 و یا ژنهای اپران لاکتوز E.coli ارائه می دهد، مشاهده نمائید. اغلب حامل های تجاری از ساختارهای Ecoli و پلازمید طبیعی اولیه مشابه مشتق می شوند. بنابراین توالی های آنها باید ۱۰۰درصد مشابه باشند. تطابقات در Univec با همان ترتیبی که در بانک اطلاعاتی اتفاق می افتند، گزارش می شوند. اگر در ترتیب مذکور حامل شما اول نیست، ورودی دیگری (مرتبط با توالی ۱۰۰ درصد مشابه) گزارش می شود.

اگر روی ژن هایی کار می کنید که مشابه ژن های مورد استفاده برای ساخت حامل هستند (مانند ژنهای مقاومت به آنتی بیوتیک، ژنهای فعال کننده/ بازدارنده و یا تعدادی از ژنهای بیوسنتز شده)، در حقیقت باید انتظار ایجاد تطابقات در بانک اطلاعاتی Univec را داشته باشید.

قبل از تفسیر کامل جزئیات alignment، این توالی ها را دور نیاندازید.



شکل ۲-۵: خروجی VecScreen نمایش داده شده از یک توالی آلوده

در هر مورد، انتظار مشاهده ۱۰۰ درصد تطابق با هر قطعه ای از ژنوم مخمر و یا *E. coli* را داشته باشید، حتی اگر این قطعات متعلق به حامل ها نباشند. این ها از علائم قطعی آلودگی هستند. تکنیکهای مدرن بیولوژی مولکولی در تکثیر قطعاتی از DNA همکار شما بسیار توانا هستند!

پردازش یا تأیید یک Restriction Map

در این مرحله، راه دیگری برای چک کردن اینکه توالی بازگشت داده شده از آزمایشگاه یا شرکت توالی یابی واقعاً با قطعه ای از DNA که شما آن را فرستاده اید، مطابقت دارد: الگوی ناشی از هضم با آنزیم محدود کننده (که به آن Restriction map می گویند) را که در عمل به مشاهده کرده اید، با الگوی تئوری restriction map که از روی توالی به دست آمده مقایسه کنید.

به دست آوردن restriction map به صورت تئوری کار ساده ای است. کافی است به تطابقات قطعی محل های شناسایی آنزیم محدود کننده در توالی تان نگاهی بیاندازید. تمامی آنزیمهای شناخته شده و محل های اتصال در بانک اطلاعاتی REBASE (rebase.neb.com) بطور کامل شرح داده شده است.

پس از یافتن محل انطباقات در توالی، کافی است تعداد قطعات حاصل نموده (قطعات DNA در بین دو محل اتصال) را شمارش و اندازه ی آنها را محاسبه کنید. در نهایت می توانید آنچه را که به صورت تئوری به دست آورده اید، با نتیجه

آزمایش واقعی خود مقایسه نمائید. هر چه آزمایش خود را با آنزیمهای محدود کننده بیشتری انجام دهید، از صحت توالی خود بیشتر مطمئن خواهید شد.

این عملیات را می توانید برای بررسی توالی های بلند، مانند توالی های ژنوم باکتری، که از تجمع تعداد زیادی قطعات کوچک تر ایجاد شده اند، نیز انجام دهید. پروتکل Shotgun DNA sequencing همین کار را انجام می دهد. اگر توالی و نحوه تجمع هر دو صحیح باشند، تعداد و اندازه قطعات پیش بینی شده نیز باید با restriction map به دست آمده از آزمایش مطابقت داشته باشند.

چند سایت اینترنتی خوب، کار جمع آوری اطلاعات در مورد سایت های آنزیم های محدود کننده و انجام تطابق را به جای شما انجام می دهند. این سایت ها نسخه های رایج از بانک های اطلاعاتی آنزیم های محدود کننده (REBASE[1]) را جمع آوری کرده اند، پس نگران نباشید. آن ها همچنین توالی شما را برای همه سایت های اختصاصی مرتبط با یک آنزیم که شما می توانید بر اساس معیار های خاصی انتخاب کنید، اسکن می کنند.

یکی از این سایت های اینترنت از ابزار فعل وانفعالی بسیار جالبی به نام Webcutter استفاده می کند و طراحی آن توسط MAX Heiman صورت گرفته است. برای کار با این سایت دستورالعمل زیر را دنبال کنید:

۱. به آدرس اینترنتی www.Firstmarket.com/cutter/cut2.html وارد شوید.
این سایت امکان دسترسی به جدید ترین نسخه webcutter را فراهم می کند.
۲. توالی را که می خواهید چک شود، انتخاب نمائید.

در این مرحله، Webcutter سه راه برای ورود توالی (بارگزاری) پیش روی شما قرار می دهد:

- آن را به داخل پنجره Sequence نمایش داده شده در صفحه ی www.firstmarket.com/cutter/cut2.html وارد نمائید.
- این توالی را از کامپیوتر خود به شبکه بارگذاری (Upload) نمائید. (استفاده کنندگان از سرور Netscape). توالی موردنظر باید به فرمت text بدون نام آن یا عنوان به سبک FASTA باشد.
- به طور مستقیم از طریق شماره دسترسی و یا روشهای شناسائی دیگر وارد یک ورودی GenBank شوید. (مورد آخر فقط زمانیکه می خواهید یک آزمایش کلونینگ برای یک ژن شناخته شده انجام بدهید، لازم است.)

۳. گزینه های مرتبط با مشکلتان را انتخاب نمائید.
این گزینه ها عبارتند از:

- تصمیم گیری در مورد اینکه توالی مورد نظر را به صورت خطی در نظر بگیریم یا حلقوی
- انتخاب فرمت های گوناگونی برای خروجی داده ها (بصورت نقشه و یا جدول آنزیم ها)
- جدا کردن نتایج بر اساس آنکه آیا آنزیم ها قادر به برش توالی شما هستند و یا خیر، یا توالی شما را به دفعات مشخصی برش می دهند یا خیر.
- مشخص کردن زیر مجموعه ای از آنزیم ها، بر اساس ملاک های گوناگون، برای شرکت در آنالیز

۴. بر روی گزینه Analyze Sequence در پائین فرم ارائه شده کلیک نمائید.
نتایج کار، به صورت کاملاً واضح و بدون نیاز به توضیح، ظاهر می شود.

[1] Restriction Enzyme Database

دانشگاه ماساچوست، واحد پزشکی، سرویس مشابه، اما با ظاهری متفاوت، در آدرس اینترنتی biotools.umassmed.edu/tacg/WWWtacg.php ارائه می دهد.

برای داشتن یک لیست از سرور های ارائه دهنده ی Restriction map بر روی لینک REBASE Related Sites در صفحه ی اصلی REBASE (rebase.neb.com) کلیک نمایید. تمامی این سرویس ها ا زلحاظ صحت یکسانند. این سرور ها را ببینید تا سروری با فرمت خروجی و ورودی دلخواهتان پیدا نمائید.

طراحی پرایمرهای PCR

واکنش زنجیره ای پلی مراز[1] (PCR) یک تکنیک آزمایشگاهی رایج برای تکثیر قطعات DNA است.

PCR چیست؟

راه اندازی PCR شامل مخلوط نمودن DNA الگو (که شامل توالی که می خواهید تکثیر کنید)، پرایمرها، ترکیبی از نوکلئوتیدها و دیگر ترکیبات بیوشیمیائی و یک آنزیم مقاوم به حرارت به نام DNA پلی مراز، در یک تیوب پلاستیکی کوچک می باشد. سپس تیوب را درون دستگاه گرانقیمت آزمایشگاهی به نام Thermal cycler قرار می دهید. بر اساس برنامه ریزی انجام شده، دستگاه دمای محتویات تیوب را بالا و پائین می برد. در هر گردش حرارتی، تعداد کپی های قطعه DNA مورد نظر دو برابر می شود.

برای مثال ، بعد از ۳۰ سیکل حرارتی ، شما 2^{30} (۱ میلیارد) نسخه جدید خواهید داشت. دقیقاً به همین علت است که PCR تکنیکی عالی در پزشکی قانونی است.

اگر می خواهید در مورد PCR بیشتر بدانید، به آدرس اینترنتی nature.umesci.maine.edu/forensics/p_intro.htm مراجعه کنید. یک دوره آموزشی به صورت انیمیشن در این آدرس خواهید دید.

برای انجام PCR باید:

۱. توالی DNA را که می خواهید تکثیر نمائید، شناسائی کنید.
۲. پرایمرها را از یک شرکت سنتز DNA سفارش دهید.

پرایمرها قطعات کوچکی از DNA (۲۰ الی ۳۰ باز) می باشند که به ابتدا و انتهای توالی موردنظر شما اتصال پیدا می کنند. طراحی پرایمر مناسب حساس ترین مرحله از PCR است.

۳. آزمایش PCR را انجام دهید.

[1] polymerase chain reaction

این روزها، محققین PCR را برای انواع آنالیز های توالی ها در آزمایشگاه استفاده می کنند، برای مثال: کلونینگ یک ژن در حامل بیان[1]، تایید نحوه سرهم شدن و توالی ژن، ایجاد کردن یا تشخیص جهش ها و یا شناسائی خویشاون های یک ژن دردیگر ارگانیسم ها.

حرفه ای ترین و در عین حال مرموزترین مرحله از PCR طراحی پرایمر است. پرایمر ها دو قطعه کوچک از DNA هستندکه می توانند در دو طرف به صورت خیلی اختصاصی متصل شوند. برنامه های طراحی پرایمر به شما کمک می کنند که تصمیم بگیرید کدام قسمت از توالی DNA بلند شما می توانند بهترین پرایمر باشند، تا از بسیاری از مشکلات احتمالی جلوگیری نماید.

سایت اینترنتی دانشگاه پزشکی ماساچوست (biotools.umassmed.edu/) لینکی به یک ابزار ساده و بسیار کامل برای طراحی پرایمر در اختیارتان می گذارد. این آدرس در واقع نسخه اینترنتی برنامه Primer3 است. این برنامه معروف توسط Steve Rosen و Helen skaletsk ساخته شده است. به منظور استفاده از آن مراحل زیر را دنبال کنید:

۱. به آدرس اینترنتی biotools.umassmed.edu/ مراجعه کنید.

صفحه Bioinformatics Resources دانشگاه علوم پزشکی ماساچوست ظاهر خواهد شد، که دسترسی به برنامه های آنالیز توالی بسیاری را برای شما ممکن می سازد. شما می توانید بعداً این برنامه ها را امتحان نمائید.

۲. در بخش DNA Sequence Analysis بر روی لینک Primer Selection Tool کلیک نمائید.

صفحه ی برنامه Primer 3 متعاقباً ظاهر می شود. این صفحه ممکن است در ظاهر ترسناک به نظر برسد (شکل ۳-۵)، ترسی به خود راه ندهید! در بسیاری از موارد می توانید بدون تغییر گزینه ها کار خود را انجام دهید.

۳. توالی خود را در پنجره Sequence وارد نمائید.

۴. بر روی گزینه Pick primers کلیک نمایید.

نتایج به سرعت ظاهر خواهد شد. این نتایج شامل نقشه ای از بهترین جفت الیگونوکلئوتیدها (پرایمر های چپ و راست) بر اساس گزینه های انتخاب شده در فرم اولیه هستند. بعلاوه ۴ پارامتر دیگر نیز پیشنهاد شده است. مکان ، طول، دمای ذوب پیش بینی شده و درصد بازهای C+G (درصد نوکلئوتیدهای گوانین و سیتوزین) برای هر الیگونوکلئوتید لیست شده، آورده شده است.

اگر آماده ایجاد تغییر در فرم پیش فرض اولیه هستید، فرم مربوطه امکان اعمال تمام شرایط آزمایشگاهی معقول و تمامی محدودیت ها را در حین پروسه انتخاب پرایمر فراهم نموده است:

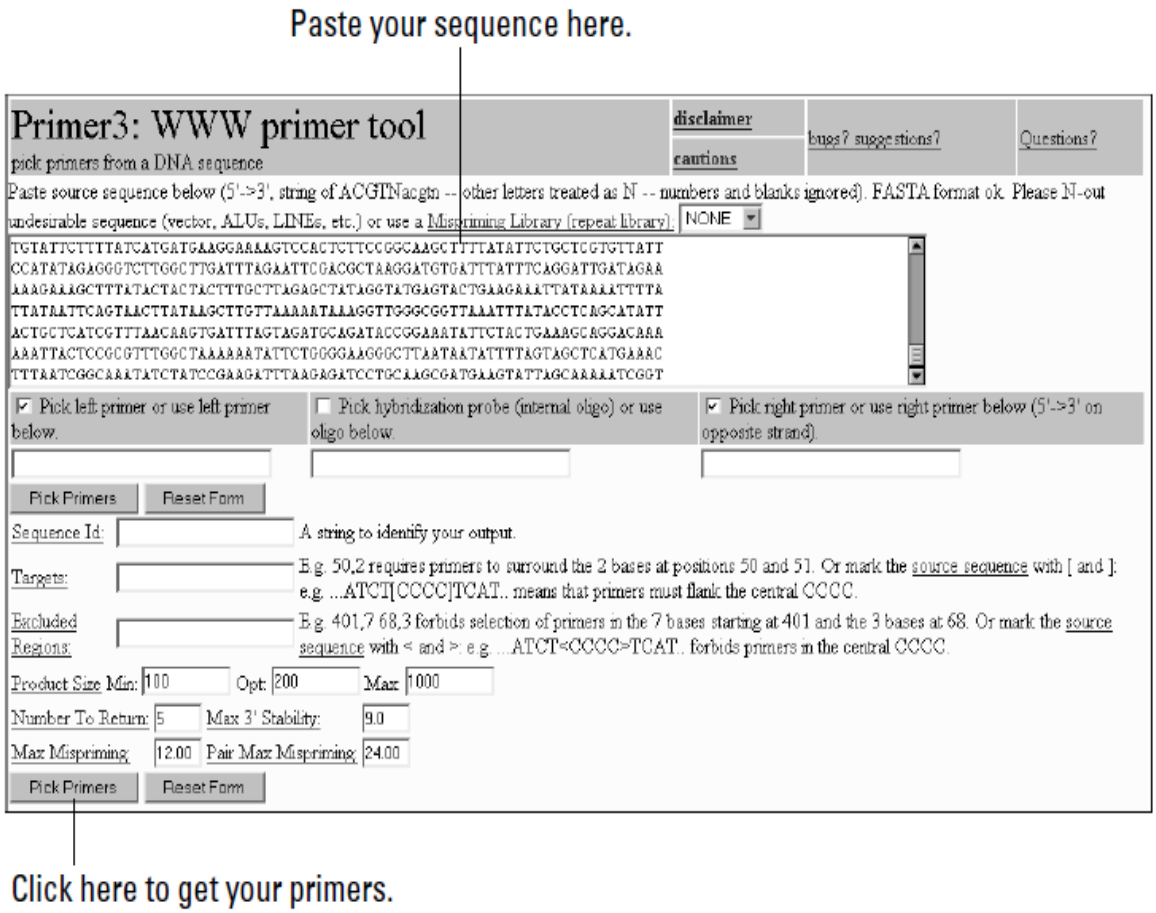
- ✓ جستجو برای فقط پرایمر راست و یا چپ، یا فقط یک پروب[2] هیبریدیزاسیون تنها.
- ✓ پیشنهاد پرایمر چپ یا راست توسط خودتان
- ✓ انتخاب مکانهایی از توالی برای اینکه در کناره ها قرار گیرند یا حذف شوند
- ✓ انتخاب یک محدوده از اندازه های محصولات
- ✓ تعیین محدوده اندازه الیگونوکلئوتید ها، درصد G+C و یا نقطه ی ذوب
- ✓ و گزینه های پیچیده دیگر

پس از دستیابی به پرایمر موردنظر، اقدام بعدی اطمینان از این مسئله است که پرایمر در نقاط دیگری، به جز در ناحیه ای که از آن ها انتظار می رود، هیبرید نشوند. مثلاً لازم است که اطمینان پیدا کنید از اینکه توالی های الیگونوکلئوتیدی انتخابی

[1] expression vector
[2] probe

شما خارج از ژن موردنظر نمی باشد و یا بدتر از این، شبیه یک توالی تکرار شونده در مولکول DNA که می خواهید تکثیر کنید، نباشد.

تکنیک های مورد استفاده برای جلوگیری از ایجاد این قبیل مشکلات شامل جستجوی توالی های حامل، همچنین ژنوم های مرتبط و و نیز توالی های تکراری در BLAST می باشد. شما می توانید از سرور NCBI BLAST برای این منظور استفاده نمائید (برای آگاهی بیشتر از جستجوهای BIAST به فصل ۷ مراجعه کنید).



شکل ۳-۵: فرم وارد نمودن اطلاعات Primer 3

آنالیز ساختار و ترکیبات DNA

پس از انجام تمامی آنالیزهای نام برده از ابتدا تا انتها، اکنون کاملاً متقاعد شده اید که توالی موجود در لوله آزمایش شما همان توالی مورد نظر است. اینک زمان آنالیز رسیده است. زمانیکه یک قطعه جدید از یک ژنوم جدید در اختیارتان است اولین سؤال دیگران از شما اینست که: درصد G+C آن چقدر است؟

علت این سوال واضح، جفت شدن نوکلئوتیدهای گوانوزین (G) و سیتوزین (C) از جفت شدن نوکلئوتیدهای آدنوزین (A) و تیمیدین (T) در نردبان DNA، محکم تر است؛ بنابر این درصد این دو نوکلئوتید نحوه ی رفتار مولکول DNA را در آزمایشات به شما نشان خواهد داد. اولین مرحله از آنالیز اینست که شما تعداد آدنوزین و گرانوزین، سیتوزین و تیمیدین را در توالی مربوطه شمارش نمائید.

تخمین محتوای G+C در توالی

اغلب اوقات، محققان محتوای G+C موجود در یک توالی را از طریق نرم افزارهای نصب شده در کامپیوتر تخمین می زنند. بسیاری از نرم افزارهای مورد استفاده در بسته بندی های تجاری هستند و در یک کنسرسیوم منبع (مانند EMBOSS) در دسترس هستند و یا توسط دانشجویان رشته کامپیوتر، البته با خرج یک پیتزا !، می توانند نوشته شوند. اگر می خواهید مجبور به خریدن پیتزا نشوید، می توانید اصول پایه زبان های برنامه نوسی کامپیوتر را فرا گیرید. با این وجود در صورتیکه تمایل به استفاده از برنامه ای آنلاین برای تخمین محتوای G+C توالی موردنظرتان را دارید، به سرور Pasteur Institute's EMBOSS (bioweb.Pasteur.fr/seqanal/EMBOSS/) سری بزنید، و در جدول برنامه های در دسترس EmBoss (bioweb.pasteur.fr/seqanal/interfaces/geecee.html)، بر روی گزینه geecee کلیک نمائید.

شمارش حروف در توالی DNA

پس از تخمین درصد G+C، مرحله پیچیده بعدی این است که مولکول DNA خود را به عنوان یک متن متشکل از کلماتی که با هم همپوشانی دارند در نظر بگیرید. برای مثال توالی DNA زیر را در نظر بگیرید:

ATGGCTGACTGACTGACTGACTGAC.....

شما می توانید آن را اینگونه بخوانید:

A, T, G, G, C, T, G, A, C, T, G, A, C, T, G, A,.....

هنگامیکه هدف شما شمارش اجزای موجود است به راحتی می توانید میزان نوکلئوتیدهای T,G,C,A را مطابق الگوی بالا شمارش نمائید. اما می توانیم توالی را به شلک زیر نیز بخوانیم:

AT, TG, GG, GC, CT, TG, GA, AC, CT, TG,....

و درصد فراوانی ۱۶ دی نوکلئوتید(کلمات دو حرفی) متفاوت را تخمین بزنید.

این بار همان توالی را به صورت کلمات سه حرفی بخوانید:

ATG, TGG, GGC, GCT, CTG, TGA, GAC, ACT, CTG, ...

و فراوانی ۶۴ تری نوکلئوتید (کلمه سه حرفی) را تخمین بزنید و همین طور ادامه دهید.

شرکت آلمانی Genomatix، وب سایت رایگان زیبائی را در اختیارتان قرار داده است. در این وب سایت می توانید فراوانی های نوکلئوتید، دی نوکلئوتید و تری نوکلئوتید توالی DNA موردنظرتان را محاسبه کنید. روش کار با آن به این شرح است:

۱. توالی DNA خود به صورت فایلی با فرمت text آماده کنید.
۲. به آدرس اینترنتی www.genomatix.de/cgi-bin/tools/tools.pl مراجعه نمائید.
۳. بر روی گزینه Create Sequence Statistics کلیک نمائید.

۴. بر روی گزینه Start Selected Task در قسمت پائین و سمت چپ صفحه کلیک نمائید. صفحه بارگذاری توالی ظاهر می شود.

۵. توالی موردنظر در پنجره ی Sequence Input وارد نمائید.

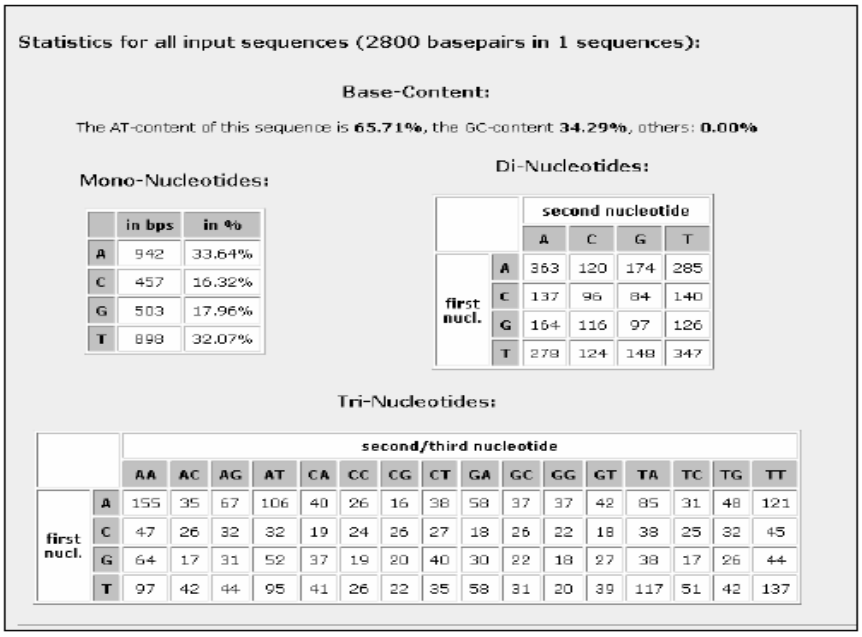
فرمت انتخابی خود را استفاده نمائید (ما در اینجا فرمت ***plain*** را استفاده می کنیم که یک نوع فرمت خام است).

۶. بر روی دکمه Load Sequence کلیک کنید.

فرم جدیدی ظاهر می گردد.

۷. فرم باز شده بدون تغییر باقی بگذارید و بر روی گزینه Start Task کلیک نمائید.

گزینه های پیش فرض درصدAT و GC، و همچنین فراوانی نوکلئوتیدها، دی نوکلئوتیدها و تری نوکلئوتیدها را محاسبه می کنند، و نتیجه را ارائه می کنند (شکل ۴-۵).



Statistics for all input sequences (2800 basepairs in 1 sequences):

Base-Content:

The AT-content of this sequence is **65.71%**, the GC-content **34.29%**, others: **0.00%**

Mono-Nucleotides:

	in bps	in %
A	942	33.64%
C	457	16.32%
G	503	17.96%
T	898	32.07%

Di-Nucleotides:

		second nucleotide			
		A	C	G	T
first nucl.	A	363	120	174	285
	C	137	96	84	140
	G	164	116	97	126
	T	278	124	148	347

Tri-Nucleotides:

		second/third nucleotide														
		AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
first nucl.	A	155	35	67	106	40	26	16	38	58	37	37	42	85	31	48	121
	C	47	26	32	32	19	24	26	27	18	26	22	18	38	25	32	45
	G	64	17	31	52	37	19	20	40	30	22	18	27	38	17	26	44
	T	97	42	44	95	41	26	22	35	58	31	20	39	117	51	42	137

شکل ۴-۵: شماتیک نتیجه ی آنالیز ساختاری

شمارش کلمات بلند تر در توالی های DNA

هنگام آنالیز مولکول DNA، شاید جالب باشد که کدهای بیش از سه نوکلئوتید را شمارش نمائید. اگرچه بسیار بعید است که دستجاتی با بیش از ۶ یا ۸ حرف استفاده شود. اگر اندازه ی آنها بیش از سه حرف باشد به آن k-tuples (K تائی)گفته می شود (مثلاً کلمه ۶ حرفی = هگزامر = tuple-). بر اساس احتمال قرارگیری حروف کنار هم دیگر ۶۴ نوع ۳ تائی، ۲۵۶ نوع ۴تائی، ۱۰۲۴ نوع ۵ تائی و ۴۵۹۶ نوع ۶تائی مختلف وجود دارد. منطقی به نظر نمی رسد که آمار هگزامرها را در یک توالی کوتاه برآورد نمائید، چون ممکن است خیلی از هگزامر ها حد اکثر یک بار دیده شوند. از طرف دیگر تخمین هگزامرها (۶ تائی ها، با فراوانی کم) با فراوانی بالای غیر قابل انتظار در یک سری از توالی ها (مثل مناطق پروموتور)، نقطه ی شروعی برای کشف بخش های تکرار شونده مناطق تنظیمی توالی است.

سرور EMBOSS در مؤسسه پاستور در پاریس، یک نسخه آنلاین برنامه ی شمارش کد ارائه می کند، که با استفاده از آن می توانید فراوانی حروف در اندازه های مختلف در توالی DNA مورد نظرتان را برآورد کنید (ما آن را تا ۲۰ حرف هم امتحان کرده ایم). در ادامه با نحوه کار این برنامه بیشتر آشنا خواهید شد:

شمارش سه تائی در برابر فراوانی کدون ها

در این بخش به توضیحی پیرامون مفهوم شمارش نوکلئوتیدهای ۳تایی (Triplets) می پردازیم این پردازش عمدتاً از طریق خواندن توالی DNA در یک رشته به حالت overlap انجام می گیرد. این اقدام نباید با محاسبه استفاده از کدون ها[1] اشتباه گرفته شود (به یاد آورید که کدون، نوکلئوتیدهای سه تائی در محتوای ژنتیکی می باشند که در هنگام ترجمه توالی DNA به پروتئین مورد استفاده قرار می گیرند، برای توضیح بیشتر به بخش ۱ مراجعه کنید). مفهوم کاربرد کدون ها فقط در رابطه با توالی های DNA حاوی نواحی کدکننده پروتئین (به عنوان مثال، قالب باز خواندن یا ORF) می باشند. بنابراین شامل محاسبه فراوانی تری نوکلئوتید ها با حالت غیر overlap و با تقلید از انجام پروسه ترجمه پروتئین می باشد. از این رو، برای یک رشته ی DNA مانند:

ATGTTTAGTGATGGACGCCCAGCATGCGAC....

فراوانی کدون ها از طریق فرایند زیر صورت می گیرد:

ATG, TTT, AGT, GAT, GGA, CGC, CCA, GCA, TGC, GAC, ...

جدول حاصل از نحوه خواندن کدون ها مشخص می کند که کدام سری از کدون ها ترجیحاً برای کد کردن آمینواسید ها استفاده می شوند. به علاوه، مشخص می کند که کدام یک به ندرت مورد استفاده قرار می گیرند. سری کدون های مورد استفاده در گونه های مختلف متفاوتند، برای مثال سری کدون های مورد استفاده در ژنهای انسانی نسبت به سری کدون های مورد استفاده در ژنهای E. coli متفاوت اند. به همین دلیل بعضی از ژنهای انسانی در این باکتری ها زیاد کارا نیستند. برای امتحان کردن یک برنامه ایجاد کننده سری کدون ها، بد نیست یک سفر مجازی به ژاپن داشته باشید، برای این منظوربه آدرس اینترنتی زیر مراجعه کنید:

www.kazusa.or.jp/codon/countcodon.html

۱. به آدرس اینترنتی bioweb.pasteur.fr مراجعه کنید.
بر روی نشانه ی English/American (در بالا و سمت راست) برای تغییر به حالت انگلیسی کلیک نمایید.
۲. بر روی لینک EMBOSS در قسمت پائین و سمت چپ صفحه کلیک کنید. صفحه مربوطه ظاهر خواهد شد.
۳. بر روی لینک wordcount کلیک نمائید.
صفحه ی زیبای wordcount ظاهر خواهد شد.
۴. آدرس پست الکترونیک خود را وارد کرده و توالی مورد نظرتان را (فرمت خام هم مورد قبول است) در فیلد مربوطه منتقل نمائید. اندازه کلمات مورد نظر را به دلخواه تنظیم کنید.
۵. بر روی گزینه Run wordcount در قسمت بالا و سمت چپ صفحه کلیک نمائید.
پس از لحظاتی، صفحه نتایج ظاهر می گردد.
۶. بر روی لینک word count.out برای مشاهده ی نتایج در قالب یک لیست ساده کلیک نمائید.

[1] codon usage

کار کردن با دیگر آنالیزهای ترکیبی DNA

به هنگام حضور در سایت موسسه پاستور، فرصت را برای کاوش بیشتر در لیست مدلهای موجود EMBOSS غنیمت بشمارید؛ برای مثال chips (برای آنالیزهای آماری سری کدون ها) و یا برنامه *cpgreport* که مناطق غنی از CpG را پیدا می کند.

یافتن تکرارهای داخلی در یک توالی

یکی از آنالیزهای ساختاری سودمند شامل مکان یابی قطعاتی است که بیش از یکبار در طول توالی تکرار شده است. به این قطعات، توالی های تکراری[1] گفته می شود.

در حقیقت میان توالی های تکراری و کلمات بلند (۶تائی، ۸تائی و یا بلند تر) تفاوت چندانی وجود ندارد. در اینجا به این قانون می رسیم که: یک تکرار، کلمه ای است که به اندازه ای بلند است که به صورت تصادفی احتمال وقوع آن در توالی بسیار کم است. برای مثال کلمه سه تائی GTC که ۴ مرتبه در طول یک توالی طولانی ۵۰۰ نوکلئوتیدی تکرار شده باشد، به عنوان یک توالی تکراری به حساب نمی آید.

تفاوت دیگر میان شمارش کلمات و آنالیز تکرارها این است که تکرار ها می توانند ناقص باشند. از طرف دیگر توالی های تکراری، بر خلاف کلمات، نیازی نیست که حتماً مشابه باشند. زیست شناسان در پی شناسائی تکرارهای پشت سر هم[2] (توالی های مشابه پشت سر هم که در یک رشته ی DNA قرار دارند) از تکرارهای معکوس[3] می باشند (توالی های مشابهی که در دو زنجیره ی مکمل هم اتفاق می افتد). زیست شناسان اغلب به توالی های تکراری علاقمند هستند، زیرا اغلب آنها در نوآرایی های ژنوم و مکانیسم های تنظیمی بیان ژن نقش دارند.

الگوریتم های گوناگون بسیاری برای یافتن توالی های تکراری در طول یک توالی DNA (و یا توالی یک پروتئین) وجود دارد. تمام این الگوریتم ها برای شناسائی قطعاتی است که به نظر نمی رسد شباهت آن ها به هم از روی شانس باشد. بخش مهم آن نمره دهی و درجه بندی قطعات با توالی های مشابه پشت سر هم می باشد.آیا تطابق کامل ۵ نوکلئوتید پشت سر هم کافی است تا آن را یک تکرار به حساب بیاوریم؟ و آیا تطابق ۹ نوکلئوتید از ۱۰ تا بهتر است یا ۱۲۳ نوکلئوتید از ۱۶۰ تا؟ کدام یک را زود تر از همه گزارش خواهید کرد؟ تا چه حد می خواهید تکرارهای محتمل را پیدا کنید؟

یافتن تکرارها یک کار ظریف است!

چون هیچ جواب جامعی برای سوالات پیرامون توالی های تکراری وجود ندارد، برنامه یابنده توالی های تکراری از شما می خواهند که یک آستانه برای الگوریتم درجه بندی آنها، اندازه تکرار، تعداد کپی ها، ، حالت تناوبی (فواصل بین تکرارها) و دیگر مواردی که معمولاً شمانمی دانید، تعیین کنید. تنظیمات پیش نویس موجود که توسط این برنامه ها اغلب در نظر گرفته می شود ممکن است برای توالی خاص مورد نظر شما کارآیی نداشته باشد.

در این باره، برآورد سریعی که ما از جستجوکننده های توالی های تکراری در پایگاه ها اینترنتی (با استفاده از توالی های حاوی تکراری که ما برای این هدف ساخته ایم) انجام داده ایم، به طور غیر منتظره در تمامی سرورهای موجود ناموفق بود.

---

[1] repeats
[2] tandem repeats
[3] inverted repeats

ولی این بدان معنی نیست که این سرورها ناکارآمدند، علت این است که مشکل موجود واقعاً پیچیده است. هرگز یک نتیجه ی منفی را از یک توالی یاب باور نکنید.

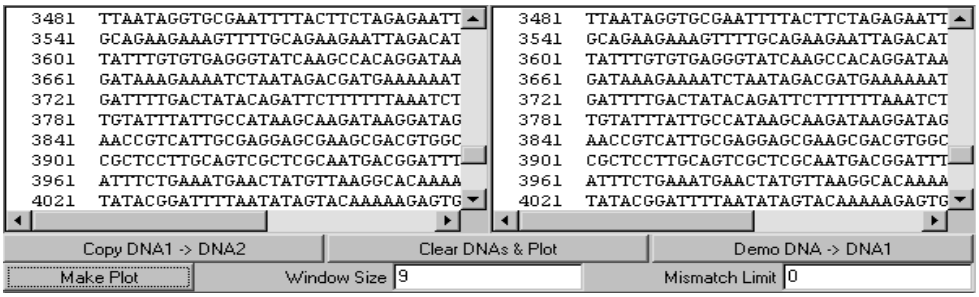
خوشبختانه، روش ساده و جهانی (DNA/پروتئین) برای حل این مشکلات، وجود دارد و آن شیوه نمودار نقطه ای[1] است،که در ادامه بصورت مرحله به مرحله نشان داده شده است (برای اطلاعات بیشتر نسبت به نمودار نقطه ای به فصل ۸ مراجعه کنید).

برای توالی های توکلئیک اسید طولانی، از برنامه dot-plot که از طریق سرویس Molecular Toolkit Web در دانشگاه Colorado در دسترس است، می توان استفاده نمود. فقط کافی است مطابق مراحل زیر عمل نمائید:

۱. به آدرس اینترنتی arbl.cvmbs.colostate.edu/molkit/ مراجعه نمائید. Molecular Toolkit مجموعه ای از برنامه هاست که به منظور آنالیز و دستکاری اطلاعات حاصل از توالی های پروتئین و نوکلئوتیک اسید بکار می رود. این برنامه تحت سیستم(Java (1.0 نوشته شده و سیستم عامل شما باید با زبان این برنامه آشنا باشد.

سایت مذکور برای اعمال دستکاری های ساده در توالی های DNA از جمله معکوس کردن رابطه مکملی زنجیره ها (جابجایی زنجیره ها)، ترجمه پروتئین و یا نمایش یک restriction map بسیار سودمند خواهد بود.

۲. بر روی لینک Dot Plots در بالای لیست کلیک نمائید. یک فرم، با دو بخش کنار هم با عناوین DNA Number1 و DNA Number2 ظاهر می گردد(شکل ۵-۵).



شکل ۵-۵: فرم نمایش داده شده از برنامه ی DNA-dot-plot.

۳. توالی خود را از یک فایل txt. و یا word کپی نمایید.
برای اینکار از گزینه copy در نوار ابزار Word یا از دکمه ی Ctrl+C به صورت همزمان استفاده نمائید.

۴. توالی خود را با استفاده از گزینه Paste در Word و یا ctrl+v در پنجره DNA Number1 وارد نمائید.
برنامه با فرمت خام توالی شما (فقط نوکلئوتید ها) مشکلی ندارد.

۵. توالی خود را به پنجره DNA Number2 نیز وارد نمائید. برای این کار می توانید از گزینه مخصوص Copy DNA1 -> DNA2 در زیر پنجره DNA Number1 استفاده نمائید.

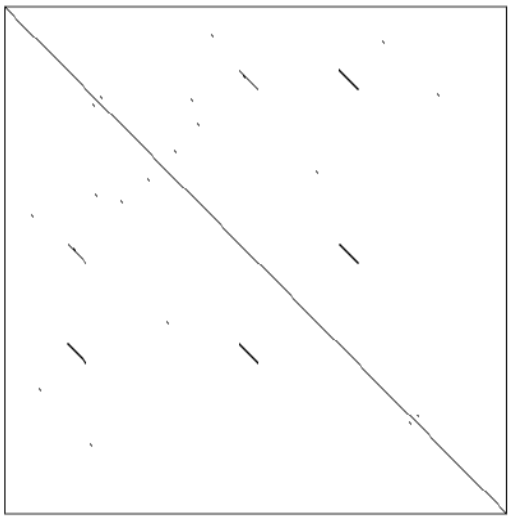
۶. بر روی گزینه Make Plot در زیر پنجره های توالی ها کلیک نمائید.

نمودار نقطه ای مربوطه ظاهر خواهد شد. به همین سادگی! اگر نمی خواهید مدت زمان زیادی را منتظر بمانید، توالی خود را به اندازه 10,000 bp محدود نمائید.

[1] dot-plot

در این مورد خاص، شما می توانید نمودار نقطه ای مربوطه را به سرعت بدست آورید. در زمانی کوتاه نمودار نقطه ای مربوطه در یک مربع در زیر پنجره ورودی نمایان خواهد شد. برای مثال در شکل ۶-۵، توالی ساخته شده است که قطعه طولانی از آن به طور ناقص سه بار تکرار شده است. این تکرار ها به صورت سه خط هم در قسمت بالائی و هم قسمت پائینی خط مورب نمودار نشان داده شده اند. با امتداد دادن آن ها تا خط مورب اصلی، مکان قطعات توالی تکراری مطابق را می توانید پیدا کنید (شکل ۶-۵). توالی های بلند در مقایسه با توالی های تکراری کم اهمیت، تصادفی و کوتاه با طول ۹ تا ۱۵ نوکلئوتید (که بصورت نقطه هایی در نمودار مشخص می شوند، کاملاً متمایز شده اند.

در صورت استفاده از توالی خود، می توانید پارامترهای Window Size (یعنی حداقل اندازه تکرار ها) و همچنین پارامترهای Mismatch Limit (یعنی تعداد نوکلئوتید های غیر مشابه) آزمایش نمائید تا ببینید که چگونه این پارامترها بر روی ظاهر نمودار شما تأثیر می گذارند.



شکل ۶-۵: نمودار نقطه ای نشاندهنده ی یک قطعه که ۳ بار تکرار شده است.

استفاده از نمودار نقطه ای برای شناسائی تکرار های معکوس

یادآوری این مطلب که برنامه ی نمودار نقطه ای قادر به جستجوی تکرارهای معکوس می باشد، بسیار حائز اهمیت است. توالی های تکراری معکوس در حقیقت مشابهت های ناحیه ای بین رشته ی مستقیم و معکوس یک توالی DNA می باشد. برای این منظور، کافی است توالی اصلی را در محور X (DNA Number1) و توالی زنجیره ی مکمل را در محور Y (DNA Number2) قرار دهید. برای ساختن یک توالی معکوس، در آدرس اینترنتی arbl.cvmbs.colostate.edu/molkit/ گزینه ی Manipulate Sequences را انتخاب نمائید. تنها تفاوت این کار با پروتکل قبلی این است که خط مورب ناپدید می گردد. تکرارهای معکوس بصورت قطعات خارج از خط قطری نمایش داده می شوند.

شناسائی تکرارهای مختص ژنوم در یک توالی

کشف یک تکرار جدید در یک توالی که لیست تکرار های آن قبلاً تثبیت شده بوده است (مثلاً خانواده تکرار های Alu در ژنوم انسان)، نباید شما را سردرگم کند. کشف یک تکرار جدید کاری است که باید با ساختار داخلی توالی خود انجام دهید. شناسائی یک تکرار تأئید شده بستگی دارد به شناسائی بعضی از مشابهت های ناحیه ای بین توالی شما و یک بانک اطلاعاتی از پیش تعیین شده از تکرار ها به عنوان مرجع است.

عناصر تکراری اغلب در ارگانیسم های چند سلولی اتفاق می افتد. ژنوم مهره داران به طور اختصاصی غنی از بسیاری از عوامل تکرار شده در اندازه های مختلف است. شما می توانید درک کاملی از موضوع را از طریق سایت تحت وب Repbase از موسسه ی تحقیقاتی اطلاعات ژنتیکی در آدرس اینترنتی www.girinst.org پیدا نمائید. این سایت امکان دسترسی به نرم افزار CENSOR را فراهم می کند. این نرم افزار توالی های مورد نظر را از بین یک مجموعه مرجع از تکرار ها غربال می کند، اتصال دهنده بخش ها را با استفاده از سمبول های خاص پوشش می دهد، و همچنین گزارشی از تکرار های پیدا شده ارائه می دهد.

یک سرویس مشابه در آدرس اینترنتی www.repeatmasker.org در یک سرور که توسط Institute for System Biology حمایت می شود، ارائه شده است. توجه داشته باشید که این برنامه به طور متوسط ۵۰٪ از توالی ژنوم انسانی را پوشش می دهد.

شناسائی مناطق کد کننده ی پروتئین

آماده باشید. از اینجا تفریح آغاز می شود! تا اینجا ما توالی DNA را برای ناخالصی های ممکن چک کرده ایم، restriction map های آن را بررسی کرده ایم، آنالیزهای آماری ساختارها را انجام داده ایم و تکرار های آن را شناسائی کردیم. حال زمان آن است که به یک مرحله مهیج برسیم! می توانیم بررسی نمائیم که آیا یک پروتئین از توالی نامفهوم ATGCTAG به وجود می آید و این کار در کجای توالی صورت می گیرد.

برای فهمیدن این که"کشف پروتئین" چیست، شما باید به خاطر آورید که ژنهای کد کننده ی پروتئین درمیکروب ها و ارگانیسم های پر سلولی ساختار های بسیار متفاوتی دارند. در میکروبها، هرپروتئین از یک قطعه ساده DNA (از ابتدا تا انتها) کد می شود که به آن ORF یا قالب باز خواندن گفته می شود. ولی در مورد ژنهای گیاهی و حیوانی و بخصوص مهره داران، پروتئین ها از قطعات کد کننده زیادی به نام اگزون بوجود می آیند که توسط قطعات DNA غیر کد کننده به نام اینترون جدا شده اند. این مطلب به وضوح نشان می دهد که چرا روشها و برنامه های استفاده شده برای یافتن پروتئین ها در میکروبها و یوکاریوتهای پیچیده اینقدر با هم متفاوتند. این بخش با یک استراتژی ساده به نام ORFing آغاز می شود، که چگونگی جستجو و یافتن مناطق کد کننده و پروتئین مربوطه را در توالی های DNA میکروبی و یا توالی های mRNA یوکاریوتی نشان داده می شود.

ORFing توالی DNA

یک توالی DNA برای کد کردن یک پروتئین ، باید حاوی کد آغازین برای ترجمه (معمولاً ATG) باشد و نیز در نزدیکی کدون آغاز هیچگونه کدون پایانی (معمولاً TGA,TAG.TAA) وجود نداشته باشد. مفهوم دقیق و البته مختصری از یک قالب باز خواندن همین است. با توجه به این که پروتئین ها به طور متوسط دارای طول حدود ۳۵۰ اسید آمینه می باشد (و تعداد ناچیزی از آنها طول کمتر از ۱۰۰ اسید آمینه دارند)، شما می توانید از از یک معیار اضافی (حد اقل اندازه) هم استفاده کنید، برای مثال تعیین کنید که فاصله کدون آغاز تا پایان حداقل ۳۰۰ نوکلئوتید باشد و این کاری است که OFRing انجام می دهد. شما می توانید برای امتحان ORF Finder در NCBI، مراحل زیر را دنبال کنید:

۱. به آدرس اینترنتی www.ncbi.nlm.nih.gov/gorf/gorf.html وارد شوید.

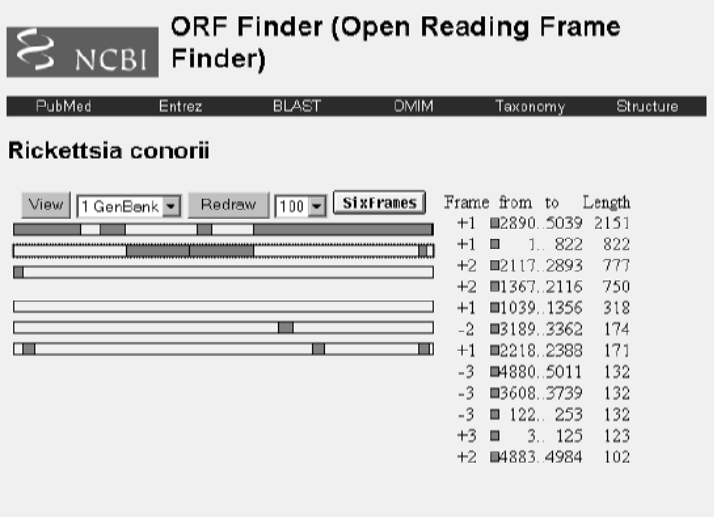
فرم ظاهر شده دارای یک جعبه ی ساده است که می توانید توالی خود را با فرمت خام آن در آنجا منتقل نمائید و نیز یک منو که کد های ژنتیکی را می توانید از آن انتخاب کنید (می توانیم شرط ببندیم که هرگز نمی دانستید که ۱۶ انتخاب برای کد های ژنتیکی وجود دارد. این تعدادبرای کد های عمومی بسیار زیاد است. اکنون اجازه بدهید کد ژنتیکی استاندارد(گزینه ی پیش نویس) را امتحان کنیم.

۲. توالی خود را از فایل با فرمت word و یا txt کپی نمائید.

ما از ۵۰۰۰ bp ابتدایی ورودیAE008509 در GenBank (ژنوم *Ricwettsia conorii*) استفاده کرده ایم. راه دیگر برای این کار اینست که شماره دسترسی را در بخش مربوطه وارد کنید و در زیر آن قسمت تعیین کنید که از نوکلئوتید ۱ تا ۵۰۰۰ را انتخاب کند. سپس به مرحله ی ۴ بروید.

۳. توالی خود را در پنجره مربوطه وارد کنید.

۴. روی گزینه OrfFind کلیک نمائید. شکل ۷-۵ یک خروجی نرمال را نشان می دهد.



شکل ۷-۵: شماتیک خروجی حاصل از برنامه جستجوگر ORF

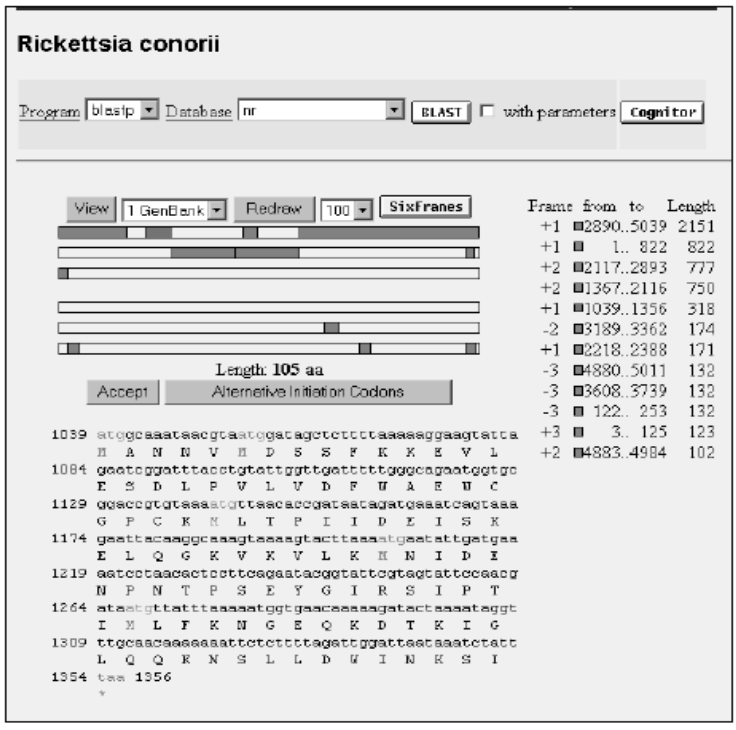
توالی DNA موردنظرتان در ۶ نوار موازی افقی، که هر نوار مرتبط با یکی از ۶ قالب محتمل ترجمه است، نشان داده شده است: ۱+، ۲+، ۳+ و بر روی رشته ی معکوس ۱-، ۲-، ۳-.

ORF های با محدوده اندازه پیش نویس (۱۰۰ نوکلئوتید) بصورت مستطیل های سبز رنگ در هر نوار افقی ظاهر می شوند. ORF ها علاوه بر این نمایش های گرافیکی، بر اساس اندازه نیز لیست شده اند. اکنون شما می توانید اندازه آستانه را تغییر دهید(۵۰، ۱۰۰، ۳۰۰).

هر چند وقت یکبار بر روی گزینه Redraw کلیک کنید خروجی حاصل تجدید شود.

به منظور بررسی دقیق تر هر کدام از ORF ها، بر روی مستطیل مربوطه در نمای گرافیکی نمایش داده شده و یا در لیست مربوطه کلیک نمائید. این اقدام موجب می گردد تا نتیجه حاصل گسترده تر شود و توالی آمینواسیدی پیش بینی شده و همچنین تعدادی گزینه تکمیلی را شامل شود (شکل ۸-۵). این گزینه ها به شما اجازه می دهد تا بعضی از بانک های اطلاعاتی NCBI را برای توالی های مشابه با این ORF، در مدت کوتاهی جستجو نمائید. این روش، ابزاری ساده و بسیار مطلوب برای ORFing یک توالی است.

قبل از آنکه به سراغ برنامه های پیچیده تر بروید، بد نیست به این نکته هم اشاره کنیم که این برنامه ی ساده همچنین قادر به شناسائی محل های کد کننده ی پروتئین در ارگانیسم های پیچیده تر نیز می باشد، البته اگر توالی شما یک cDNA باشد. cDNA (که در حقیقت شکلی از mRNA است) شامل اینترون نمی باشد و دارای ساختارهای ORF ساده و شبیه به نوع میکروبی است. بنابراین اگر می خواهید تنها مناطق کد کننده ی پروتئین درون توالی mRNA/c DNA انسانی را تعیین کنید، نیازی به استفاده از دیگر برنامه ی پیچیده نیست.



شکل ۸-۵: نتیجه ثانویه NCBI ORF Finder

آنالیز DNA از طریق GeneMark

ساده ترین روش های ORFing می توانند در حدود ۸۵ در صد از مناطق کد کننده ی پروتئین را شناسائی نماید. با این وجود موقعیت های زیادی پیش می آید که نیاز به روش های پیچیده تر(مثلاً روشی توسط GeneMark انجام می شود) به وجود می آید. این موقعیت های شامل موارد زیر می شوند:

- ✓ یافتن پروتئین های بسیار کوچک
- ✓ حل نمودن موارد مبهمی که در آن ها محل ORF های دارای همپوشانی، در قالب های باز خواندن متفاوتی (مثلاً بر روی رشته ی مستقیم و معکوس DNA) پیش بینی می شود.
- ✓ تمایل به مشخص نمودن دقیق کدون آغازین (دورترین کدون ATG همیشه کدون آغاز نیست)

معیار GeneMark برای جستجوی محل های کد کننده کمی پیچیده تر از این معیار است که:" این مناطق بایستی یک قالب خواندن بدون تداخل با مناطق دیگر و دارای یک حداقل طول معین باشد". این برنامه ویژگی های آماری توالی را نیز مد نظر قرار می دهد (بسیار شبیه به word/codon usage که قبلاً به آن اشاره شد) و انواع مختلفی ضریب های احتمالاتی کیفیت را برای هر کاندیدای ORF بودن تعیین می

کند. در حین انجام پروسه، بعضی از ORF های کوچک ممکن است تغییر رتبه (از نظر آماری) پیدا کنند و بدین ترتیب مکان دقیق آغاز تغییر کند. ضریب های کیفیت برآوردی از میزان شباهت بین ORF های کاندید شده و مدل ژنی ایده آل می باشد. مدل تکمیل شده این مدل، Markov می باشد که در حقیقت یک مفهوم ریاضی فراتر از مطالب این کتاب است. خوشبختانه استفاده از این برنامه ها از فهمیدن آن ها آسان تر است. مراحل زیر نحوه استفاده از آن ها را نشان می دهد:

۱. به آدرس اینترنتی /exon.gatech.edu/GeneMark وارد شوید.

این صفحه ی اصلی GeneMark است که از طرف Georgia tech درآتلانتا راه اندازی شده است. این صفحه شامل نسخه های تخصصی گوناگونی از این برنامه (که هر برنامه مختص مدل ژنی خاصی است) برای کاربر روی پروکاریوت ها (میکروبها)، یوکاریوتها (حیوانات)، cDNA (نسخه ی DNA بدست آمده از mRNA) و یا توالی های ویروسی است. در مورد مثالی که ما دنبال می کنیم، در این مرحله می خواهیم یک توالی از یک میکروب را که شناخت ناچیزی از آن داریم، آنالیزکنیم.

۲. در بخش Bacteria/Archeae بر روی لینک Heuristic models کلیک نمائید.

با این کار نسخه ای از برنامه که مرتبط با آنالیز توالی یک ارگانیسم جدید است، را انتخاب نموده اید. به این ترتیب فرمی برای وارد نمودن توالی ظاهر می گردد.

۳. توالی مربوطه را از یک فایل با فرمت txt. و یا word کپی کنید.

پیشنهاد می کنیم از یک توالی میکروبی با طولی در حدود ۵۰۰۰bp استفاده نمائید.

۴. توالی را به پنجره مربوطه منتقل نمائید، این برنامه، مشابه دیگر برنامه ها، با فرمت خام (که فقط شامل نوکلوتید ها می شود) سازگار است.

۵. بر روی گزینه Start GeneMark.hmm کلیک نمائید تا جستجو آغاز شود.

پس از مدتی، برنامه لیستی از ژنهای پیش بینی شده، به همراه جایگاه هر کدام، در یک فرمت ساده ارائه می کند (شکل ۹-۵).

Gene Prediction Results

Information on input sequence

```
Sequence title: Wed Sep 25 13:56:18 EDT 2002
Length:         5040 bp
G+C percentage: 31.41 %
```

Parse predicted by GeneMark.hmm 2.0

```
GeneMark.hmm PROKARYOTIC (Version 2.1)
Sequence file name: sequence,    RBS: N
Model file name: heuristic_no_rbs.mat
Model organism: Heuristic_model
Wed Sep 25 13:56:28 2002
```

Predicted genes

Gene #	Strand	LeftEnd	RightEnd	Gene Length	Class
1	+	1	822	822	1
2	+	1039	1356	318	1
3	+	1367	2116	750	1
4	+	2117	2893	777	1
5	+	2890	>5040	2151	1

شکل ۹-۵: نتیجه جستجوی GeoeMark

اگر در این فکر هستید که تفاوت مدل Markov نسبت به روش ساده ORFing چیست، جواب شما تفاوت های بین دو شکل ۹-۵ و ۸-۵ است، که هر دو مربوط به یک توالی هستند. GeneMark تنها ۵ ژن بالایی لیست نتیجه ORFing را نگه می دارد! توجه کنید که شما می توانید GeneMark (در نسخه شناخته شده) را از سایت موسسه بیوانفورماتیک اروپا[1] (www.ebi.ac.uk/genemark/) نیز اجرا نمائید.

یافتن اگزونهای داخلی در توالی های ژنومی مهره داران

به هنگام پیش بینی پروتئین از روی رشته یDNA مربوطه، مهمترین مشکل پیشرو آنالیز قطعه ی توالی DNA ژنومیک انسانی است. قبلاً به این موضوع اشاره شد که ژن های مهره داران ساختار موزاییکی دارد، و قطعات کوچک پروتئین توسط اگزون هایی که توسط اینترون ها از هم جدا شده اند، کد می شوند.زمانیکه به یک توالی ژنومی انسانی نگاه می کنید، اولین سوالی که به ذهن می رسداین است که: "آیا اگزون کد کننده ی پروتئین در آن وجود دارد یا نه؟"

در راستای کارهائی که متخصصین بیولوژی مولکولی انجام می دهند، یک اگزون کدکننده ی پروتئین در حقیقت یک ORF احاطه شده توسط دو سیگنال خاص بنام *splice sites* می باشد. برنامه های بسیاری وجود دارند که می توانند این اگزون ها را تشخیص دهند. همه ی این برنامه ها بر یک اصل کلی استوار ند: در اولین قدم، آنها ORF های کاندیدای دارای خصوصیات ساختاری مناسب (مانند فراوانی کلمات و یا codon usage) را شناسائی می کنند. این همان کاری است که GeneMark انجام می دهد. در مرحله بعد، از میان آنها فقط مواردی را که توسط spice-site مناسب احاطه شده اند، نگه می دارند.

[1] European Bioinformatics Institute

اگزون های مهره داران کوچکند (با متوسط طول bp ۱۵۰) و توالی های spice-site آن ها بسیار متغیرند. به همین علت یافتن اگزون از جستجوی ORF های یک DNA میکروبی مشکل تر است. بنابراین توقع نداشتنه باشید که یک برنامه تشخیص اگزون به خوبی برنامه هایGeneMark و یا ORF Finder کار کند (در واقع برنامه های تشخیص اگزون روی هدف های بزرگتری کار می کنند). اما ما می خواهیم کار هایی را که یک برنامه تشخیص اگزون قادر به انجام آن است نشان دهیم. مراحل زیر را که از برنامه MZEF (توسط Micheal Zhang) استفاده می کند، دنبال کنید:

۱. به آدرس اینترنتی /rulai.cshl.edu بروید.

این صفحه ی اصلی آزمایشگاه Zhang در مرکز Cold spring Horbor است.

۲. در صفحه ی بعد در بخش Software Tools ، بر روی لینک Gene-Finding کلیک نمائید.

۳. در صفحه ی Gene Finder بعدی، گزینه ی Human را انتخاب نمائید.

با این کار نسخه ای از برنامه که دقیقاً برای آمارگیری های مناطق کدکننده انسانی کالیبره شده است، انتخاب می شود. سپس فرم ساده ای ظاهر خواهد شد.

۴. توالی خود را از فایلی با فرمت txt. و یا word کپی نمائید.

اگر توالی در دسترس ندارید، می توانید توالی AF018429 را از GenBank، در سایت NCBI (http://www.ncbi.nlm.nih.gov) استخراج کنید. این توالی حاوی اگزوهای شماره ۱ و ۲ از ژن dUTPase می باشد. فرمت توالی بدست آمده باید FASTA باشد.

۵. توالی خود را در پنجره مربوطه وارد نمائید.

۶. برای شروع آنالیز بر روی گزینه Submit کلیک نمائید.

برنامه به سرعت نتیجه ای را مطابق شکل ۱۰-۵ نمایش می دهد.MZEF، یک اگزون را بطور صحیح در مکانهای ۱۰۵۶ تا ۱۱۷۲ شناسائی نموده است. ادامه ی خط نتیجه شامل ارزش های عددی نشاندهنده کیفیت پیش بینی ها، انواع قالب های خواندن (FR3, FR , FR1)، splice sites و نیز پتانسیل کد کنندگی است. پیش بینی صورت گرفته ۵۰ درصد صحیح است زیرا این برنامه یک اگزون را نادیده گرفته است. به علاوه اگزون پیش بینی شده در حقیقت از مکان ۱۰۱۸ شروع می گردد. عجب نتیجه موفقیت آمیزی! نصف نسخ بدست آمده هم به طور کامل درست نیست! این میزان موفقیت در مورد برنامه های جستجوی اگزون معمول است.

یک برنامه ی پیش بینی کننده ی اگزون، که در حقیقت MZEF را با دیگر برنامه ها ترکیب می کند و باعث افزایش کارآمدی آن می شود، در سایت Michigan Tech (genome.cs.mtu.edu/aat/aat.html) در دسترس است.

MZEF Results - human - gi|2443575|gb|AF018429.1|HSDUT1 Homo sapiens dUTPase gene, exons 1 2
Wed Sep 25 15:14:25 2002

Strand = 1
Overlap = 0
Prior Prob. = .02

Internal coding exons predicted by MZEF
Sequence_length: 1735 G+C_content: 0.610

Coordinates	P	Fr1	Fr2	Fr3	Orf	3ss	Cds	5ss
1056 - 1172	0.873	0.663	0.471	0.459	112	0.467	0.581	0.668

For more information about MZEF click here

شکل ۱۰-۵: نتیجه ی MZEF حاصل از ورودی توالی AF018429 در GeneBank

تجزیه کامل ژن برای ژنوم های یوکاریوتی

اگر می خواهید به صورت جدی توالی یابی ژنوم را انجام بدهید، به بالاترین مهارت در تفسیر توسط کامپیوتر نیاز دارید، که توسط نرم افزار های genome parsing تأمین می شود. این برنامه ها قادر به بررسی قطعات بلند (۱۰۰۰۰۰تا چندین میلیونbp) از یک توالی ژنومی و پیش بینی ساختار دقیق و جزئی اگزون/ اینترون تمامی ژنها می باشند.

مشابه با برنامه MZEF، این برنامه نیز داری ساختاری چند قسمتی است،که هر قسمت به منظور شناسائی یک جزء، مثلاً اگزون های کد کننده، اگزون اول یا آخر، نواحی پروموتور، محلهای پلی آدنیلاسیون و نظایر آن، طراحی شده است. نتایج حاصل از این قسمت های مستقل در نهایت بصورت پیش بینی های منسجم ساختار ژن با هم ترکیب می شوند (مثلاً محدود نمودن Splicing اگزون ها به قالبهای خواندن سازگار). در نهایت پیش بینی های حاصل، از طریق تطابق با مدل ژنی ایده آل درجه بندی می شوند. مدل های Markov و نیز بهینه سازی برنامه ریزی های دینامیک باعث می شوند این برنامه ها اینگونه مؤثر باشند.

جدید ترین تجزیه گر های ژنوم مانند GenomeScan اطلاعات مربوط به مشابهت توالی های پروتئین را نیز مد نظر قرار می دهند.

این برنامه ها، با وجود پیچیدگی روز افزون الگوریتم های آن ها، هنوز سادگی خود را در هنگام استفاده حفظ نموده است. تنها کاری که باید انجام دهید اینست که توالی موردنظر را به یک پنجره ورودی وارد و روی یک گزینه کلیک کنید و به همین سادگی ژن تجزیه و تحلیل خواهد شد.

آنالیز یک توالی با برنامه GenomeScan

برای آماده سازی شرایط، ابتدا توالی ژنوم مهره دار مورد نظر را به فرمت FASTA در آورید. این ژنوم باید به اندازه کافی بلند باشد (۱۰۰۰۰۰bp) تا شامل حداقل یک ژن کامل باشد.راه دیگر اینست که از یک توالی ثبت شده در سایت GenomeScan استفاده نمائید.

علاوه بر این، به توالی های پروتئین (در فرمت FASTA) نیاز دارید که تشابهات بارز با مناطق کد کننده ی بالقوه در توالی DNA شما داشته باشند این قبیل پروتین ها را، با استفاده از انجام پروسه ی مقایسه ای blastx با توالی خود، می توانید پیدا نمائید (برای مطالعه بیشتر پیرامون blastx به بخش ۷ مراجعه نمائید). راه دیگر این است که از پروتئین های موجود در سایت GenomeScan استفاده نمائید.

۱. به آدرس اینترنتی genes.mit.edu/genomescan/ بروید.

صفحه ی اصلی GenomeScan که توسط موسسه ی تکنولوژی Massachusetts فراهم شده است ظاهر می گردد. GenomeScan در حقیقت جانشین GeneScan است که توسط Chris Burger ارائه شده است. تفاوت اصلی آن ها این است که در GeneScan از اطلاعات همولوژی پروتئین ها استفاده می شود.

۲. بر روی لینک GenomeScan Webserver کلیک نمائید تا یک فرم ظاهر شود.

۳. از منوی Organism، گزینه Vertebrate را انتخاب نمائید.

با این کار نسخه ای از برنامه را که با شاخص های آماری نواحی کدکننده انسانی انطباق دارد، انتخاب می شود.

۴. توالی DNA مورد نظر را از فایلی با فرمت txt و یا word کپی نمائید.

فراموش نکنید که سر برگ فرمت FASTA را هم کپی کنید.

۵. توالی DNA موردنظر را که کپی نموده اید به پنجره DNA Sequence (بالای دو پنجره ورودی بزرگ تر)وارد نمائید.

به عنوان یک راه فرعی، می توانید بر روی لینک DNA testfile در بالای صفحه کلیک کنید تا از یک توالی آزمایشی استفاده نمائید.

۶. مجموعه توالی های همولوگ پروتئین را از یک فایل با فرمت txt. و یا word کپی نمائید.

همچنین می توانید بر روی لینک protein file در بخش بالای صفحه کلیک کنید تا از توالی پروتئین آزمایشی استفاده نمائید.

هر توالی باید سربرگ FASTA داشته باشد. هر مجموعه ممکن است شامل مجموعه ی پروتئین بسیار بلند شود که بوسیله یک سربرگ از هم جدا می شوند. فایل ها باید شامل کمتر ۱ میلیون باز باشد.

۷. توالی پروتئین مربوطه را در مربع Protein Sequence، در زیر دو پنجره ورودی بزرگتر، وارد کنید.

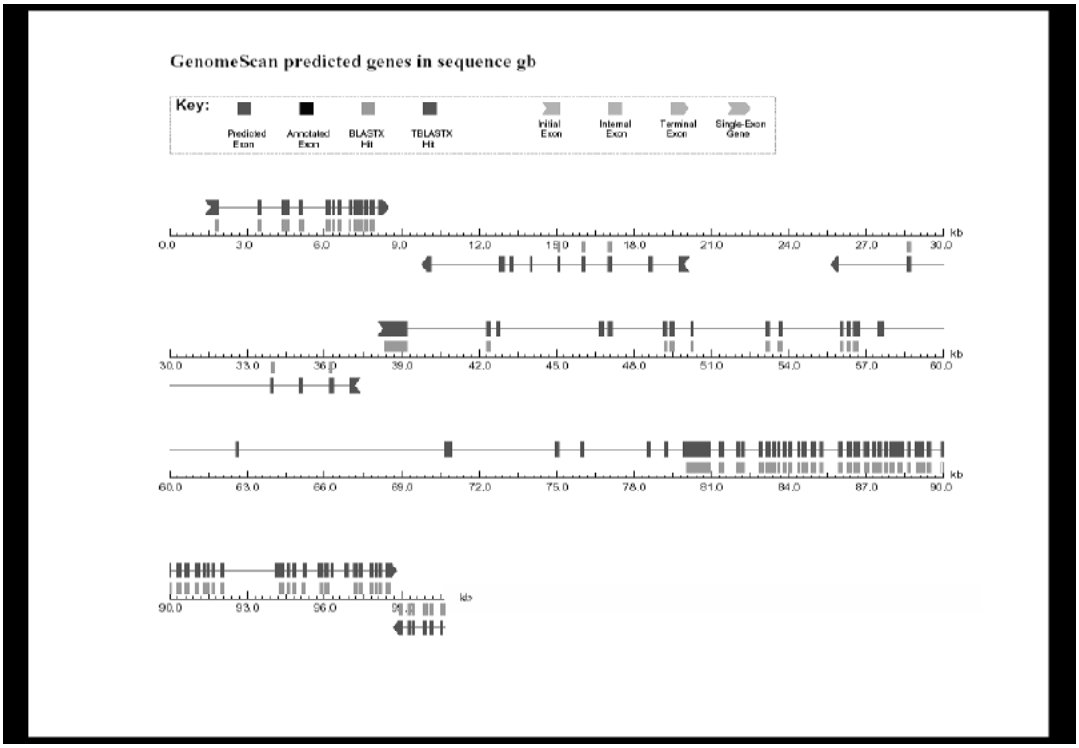
۸. به منظور آغاز عملیات آنالیز بر روی گزینه Run GenomeScan در زیر فرم کلیک نمائید.

نتیجه ی حاصل، جدول طولانی است که تمامی اجزای ژنهای گوناگون پیش بینی شده را به همراه مختصات و شاخص های کیفیت آن ها، لیست نموده است. این خروجی توسط برنامه های کامپیوتری در قالب پروژه های توالی یابی و نیز تفسیر های در مقیاس بزرگ، خوانده می شود. خوشبختانه GenomeScan قادر به ساخت خروجی گرافیکی زیبائی در بصورت PDF و تصاویر PostScripf می باشد.

شکل ۱۱-۵ نمونه ای از خروجی گرافیکی ساخته شده از توالی های DNA و پروتئین آزمایشی را نشان می دهد. (به علت آنکه تصاویر سیاه و سفید می باشند شما مجبور هستید از تصورات خود نسبت به رنگ ها استفاده نمائید که برای همین از شما پوزش می طلبیم!). ژنهای پیش بینی شده (خطوط قرمز) و اگزون های پیش بینی شده (مربع های قرمز) به وضوح نشان داده شده اند. همچنین، مدرک تأیید کننده حاصل از آنالیز تشابهات در blastx (مربع های سبز رنگ در پایین) نیز ارائه شده است. جمعاً ۵ ژن از توالی ژنومی مهره دار به طول ۱۰۰۰۰۰bp شناسائی شده است.

تجمع قطعات یک توالی

ابزارهای توالی یابی رایج، بسته به نوع تکنولوژی به کار رفته در آن ها، تنها قادر به ساخت توالی هایی به طول ۵۰۰ تا ۱۰۰۰ نوکلئوتید می باشند. بنابر این اگر توالی DNA موردنظرتان طولانی تر از این است، در نتیجه ی همپوشانی و همگذاری قطعات کوتاه بدست آمده است. تحقیق در زمینه ی الگوریتم های تجمع گوناگون و همچنین تحقیقات صورت گرفته به منظور دستیابی به نرم افزارهای تجمع بهتر، بخش حیاتی و پراهمیتی را در بیوانفورماتیک شامل می شود.



شکل ۱۱-۵: نمونه ای از خروجی گرافیکی حاصل از GenomeScan

مدیریت پروژه های توالی یابی بلند توسط نرم افزارهای معروف

در انجام پروژه ای تحت عنوان shotgun، بسیاری از محققین با کنار هم قرار دادن از بیش از ۵۰۰۰۰ قطعه ی منفرد (که به آن ها reads گفته می شود)، به یک توالی ژنوم میکروبی معمول (۴Mb) دست یافتند. برنامه های مورد استفاده برای اینگونه توالی یابی ها، یا پروژه های بزرگتر که ژنوم حیوانی /گیاهی را نیز شامل می شود، در حقیقت اطلاعات و ورودی خود را مستقیماً از کروماتوگرام ها[1] و یا نشانه ها[2]ی ابزارهای توالی یابی بدست می آورند.

این نشانه ها در حقیقت پروفایل های فلورسنت پیچیده به همراه پیک ها و دره هایی هستند که ترتیب و نوع هرنوکلئوتید(A, T, C, G) را در یک read نشان می دهند. این قبیل برنامه ها یک سیستم فراخوانی باز دارند که برای هر نوکلئوتید در هر موقعیت از یک read یک رتبه کیفیت تعیین می کند.آن ها همچنین دارای یک سیستم مدیریت داده به همراه ابزارهای نمایش و ویرایشگر متعامل هستند. به علت بسته کامل این نرم افزار های توالی یابی ژنوم به دلیل پیچیدگی شان ، در سایت های تحت web موجود نمی باشند. شما باید آنها را بر روی یک سیستم عامل قدرتمند و اختصاصی نصب نمائید. استفاده از این قبیل برنامه ها همچنین مستلزم مطالعه ی جدی است.

در ادامه کار با مشهورترین بسته های توالی یاب ژنوم که به طور گسترده ای به منظور استفاده های آکادمیک در دسترس است آشنا خواهید شد:

- ✓ **The Staden Package:** staden.sourceforge.net/
- ✓ **Phred/Phrap/Consed :** www.phrap.org/
- ✓ **ماشین Acembly که در مجموعه برنامه های معروف AceDB وجود دارد:** www.acedb.org/
- ✓ **برنامه ی Assembler تهیه شده توسط موسسه ی تحقیقاتی ژنوم (TIGR):** www.tigr.org/software/

[1] chromatogram
[2] traces

همچنین، بسته های نرم افزاری توالی یاب ژنوم بصورت تجاری از شرکت های زیادی در دسترس است از جمله:

- ✓ **Sequen cher (از شرکت Gene Codes):** www.genecodes.com/
- ✓ **Lasergene (از شرکت DNASTAR):** www.dnastar.com/

اعتقاد ما بر این است که یک پروژه طولانی توالی یابی ژنوم از مبحث این کتاب خارج است. با این وجود، هنوز ممکن است طی پژوهش های خود به روش تجمع توالی ها نیاز پیدا کنید. برای مثال، شاید در حال کار بر روی یک cDNA هستید که تنها طولی در حدود چند هزار bp دارد و توسط چند قطعه PCR ارائه شده است. شما ممکن است شماری از برچسب های توالی بیانی[1] (ESTs: cDNA های نسبی با توالی یابی قابل قبول) را پدید آورده باشید. حال می خواهید بدانید که آیا می توانید cDNA کامل و نسبی را از طریق این بر چسب ها استنباط نمائید یا خیر؟ به عنوان یک روش فرعی، شما ممکن است بخواهید EST هایی را که از یک بانک اطلاعاتی به دست آورده اید، روی هم جمع کنید. در این حالت ممکن است نشانه ها و یا درجه ی کیفیت فراخوانی باز ها را نداشته باشید.

تنها چیزی که شما در جستجوی آن هستید، برنامه ای ساده است که قادر به شناسائی همپوشانی های مهم بین قطعات شما باشد و آنها را به شکل یک توالی منفرد (که به آن Contig گفته می شود) تبدیل نماید. البته، توالی های قطعات مورد نظر ممکن است از زنجیره های DNA مختلفی بدست آمده باشد که در این حالت شناسایی چشمی همپوشانی قطعات شما را خیلی دشوار می کند. در ادامه نحوه ی انجام این پروسه را با برنامه ی تجمع توالی نشان می دهد، کافی است به سایت http://genome.cs.mtu.edu سری بزنید. ما از برنامه CAP3 که تحت یک سرور مشهور در لیون فرانسه است، استفاده خواهیم کرد.

تجمع توالی های مورد نظر از طریق CAP3

مجموعه توالی موردنظر را که می خواهید کنار هم قرار دهید، با فرمت FASTA به صورت یک فایل .txt و یا Word آماده کنید. هر توالی باید در فرمت FASTA باشد و سربرگ مختص خود را داشته باشد.

برای سادگی کار می توانید از نسخه آزمایشی مجموعه داده های تهیه شده توسط CAP3 در آدرس اینترنتی http://genome.cs.mtu.edu/cap/data استفاده نمایید. برای این منظور فایل seq (توالی) را باز نموده و ۶ توالی تحت فرمت FASTA (R1 تا R) را با استفاده از گزینه های Select All و Copy، کپی نمائید. حال همه چیز آماده است تا تمامی آنها را با استفاده از CAP3 در پایگاه بیوانفورماتیک lyon کنار یکدیگر قرار دهید:

۱. به آدرس اینترنتی http://pbil.univ-lyon1.fr/cap3.php بروید.

یک فرم حاوی یک پنجره ورودی توالی در بالای گزینه Submit و یک گزینه Clear ظاهر خواهد شد. سرور Lyon می تواند توالی های نوکلئوتیدی به طول 50 kb را پردازش نماید، که معمولاً برای هدف کنار هم قرار دادن قطعات یک cDNA کافی است.

۲. مجموعه توالی ها را به پنجره Sequence input وارد نمائید.

ما از توالی CAP3آزمایشی R1 تا R استفاده نمودیم.

[1] expression sequence tags

توجه کنید که سربرگ FASTA را فراموش ننمائید. (این تنها راهی است که برنامه می تواند یک قطعه توالی را از قطعه دیگر تشخیص دهد. هیچ مانعی در استفاده از مخلوطی توالی ها در جهت های متفاوت وجود ندارد. برنامه موردنظر به طور اتوماتیک توالی ها را با دو جهت ممکن امتحان می نماید.

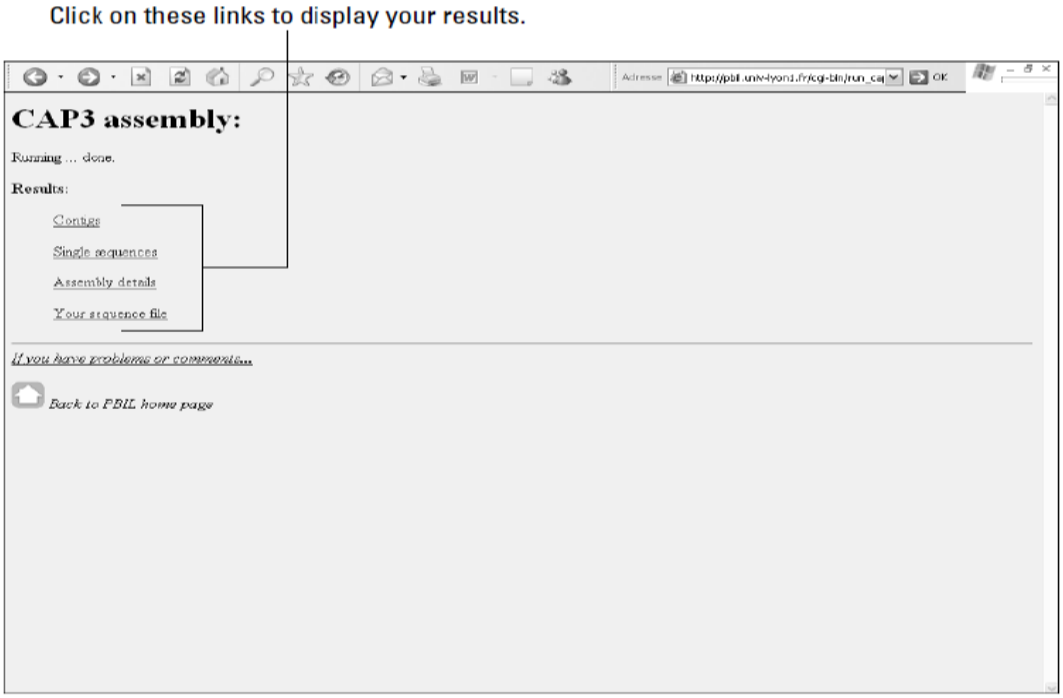
۳. گزینه SUBMIT را برای انجام عملیات تجمع قطعات انتخاب نمائید.

سرور برنامه ی CAP3 را با تنظیمات پیش نویس اجرا می کند. با استفاده از تنظیمات می توانید حداقل طول و حداقل درجه تشابه همپوشانی ها برای اینکه برنامه بتواند آن قطعات را کنار هم قرار دهد را تعیین کنید و نیز یک فایل کیفیت توالی ها را به یک فایل داده های توالی ها ارتباط دهید (برای دستیابی به جزئیات بیشتر به بخش CAP3 documentation مراجعه کنید).

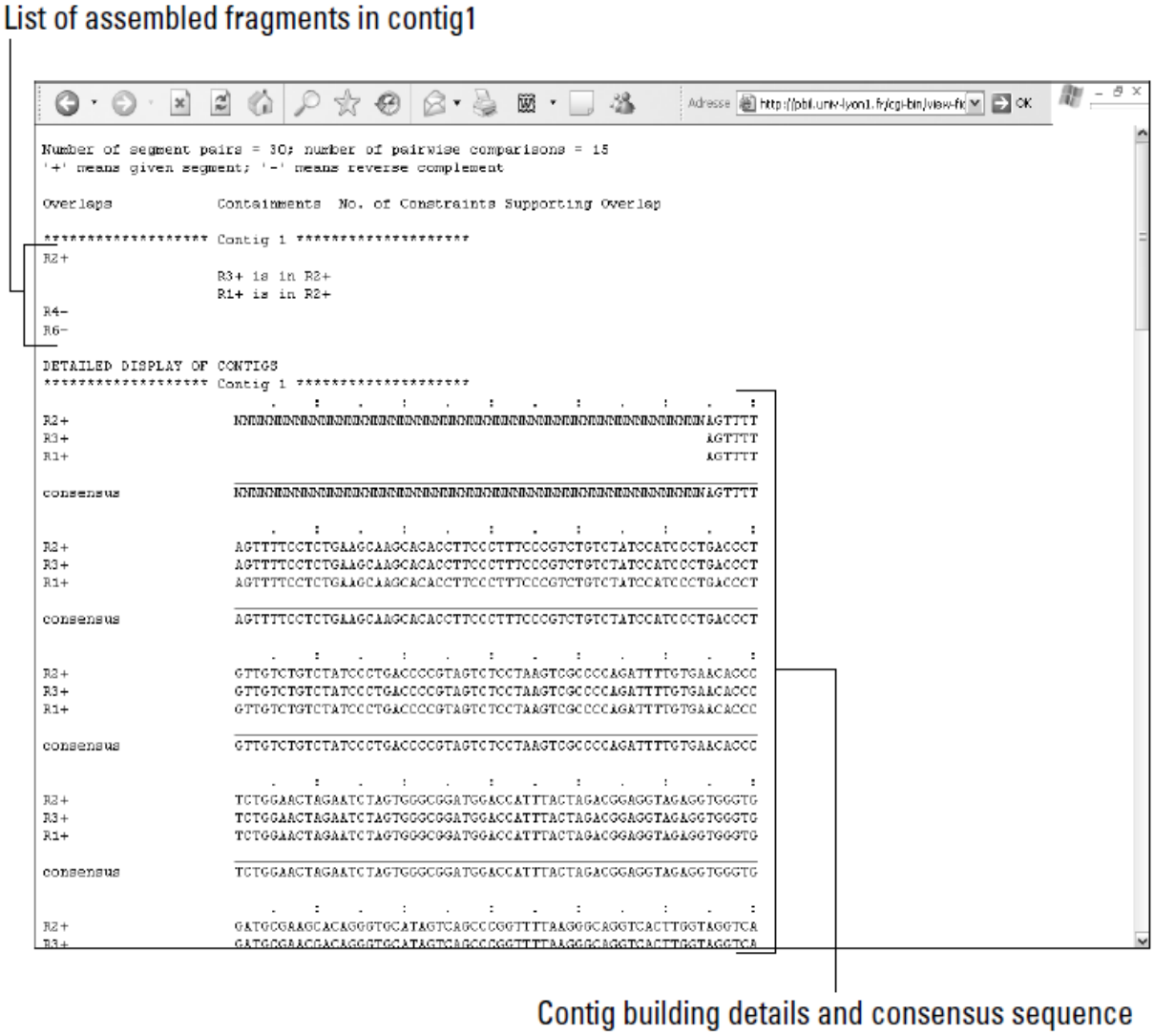
اگر از توالی های آزمایشی استفاده کنید، برنامه براحتی و به سرعت پاسخگو خواهد بود و خروجی حاصل مطابق شکل ۱۲-۵ در ۴ لینک نمایان می شود:

- ✓ **Contigs:** حاوی توالی های تجمع یافته نهایی
- ✓ **Single seqneuce:** حاوی قطعات توالی ورودی که در پروسه ی تجمع بکار گرفته نشده اند.
- ✓ **Assembly Details:** ارتباط دقیق بین قطعات موجود در Contigs
- ✓ **Your sequence File:** خلاصه ای از قطعات توالی ورودی

بر روی هر کدام از لینک ها کلیک کنید تا محتوای آن را ببینید. برای مثال، با کلیک بر روی لینک Assembly Details همپوشانی ها و ساختارهای Contig را خواهید دید (شکل ۱۳-۵). آنچه مشخص است اینست که R5 بخشی از تجمع نیست و باید در فایل Single Sequences به دنبال آن بگیرد. توالی Contig حاصل از طریق کلیک بر روی لینک Contigs نمایان می گردد (شکل ۱۲-۵).



شکل ۱۲-۵: فایل های خروجی حاصل از برنامه ی تجمع CAP3



شکل ۱۳-۵: جزئیات مربوط به همگذاری CAP3

فراتر از این فصل

این بخش پیش زمینه ی خوبی از انواع آنالیزهای گوناگون را که محققین معمولاً بروی توالی های DNA تعیین شده انجام می دهند، در اختیارتان گذاشت. البته نکات بسیار دیگری نیز به هنگام انجام پروسه های تخصصی وجود دارد که به آنها نیاز پیدا خواهید نمود. مهمترین موضوعی که در اینجا به آن پرداخته نشده است جستجو برای یافتن سیگنالها مهم بیولوژیکی در بخش غیر کدکننده توالی های DNA می باشد. این موضوع بحث ها داغی مثل پیش بینی مناطق پروموتور، عوامل تنظیم کننده و محل های اتصال پروتئین را شامل می شود.

ما به این دلیل به این موضوعات به صورت جزئی نمی پردازیم که هنوز درصد موفقیت این آنالیزها پایین است. به علاوه بهترین برنامه ها و بانک های اطلاعاتی مربوطه نیز همیشه بصورت همگانی و یا از طریق سرورهای تحت وب در دسترس نمی باشند. جدول ۱-۵ تعدادی از ورودی های مربوط به این موضوع را به شما نشان میدهد.

Table 5-1 Free Web Sites for Searching Signals in DNA Sequences

Institution; Program	*URL*	*Description*
Zhang Lab, Cold Spring Harbor, USA	`rulai.cshl.edu/`	Search for promoters and various conserved elements
Center for Information Technology, NIH, USA; WWW Signal Scan	`bimas.dcrt.nih.gov/molbio/signal/`	Search for transcriptional elements
Center for Information Technology, NIH, USA; WWW Promoter Scan	`bimas.dcrt.nih.gov/molbio/proscan/`	Predict eukaryotic promoter regions
Munich Bioinformatics Center, GER, CREDO	`mips.gsf.de/services/analysis`	*Cis*-Regulatory Element Detection
San Diego Supercomputer Center, USA; MEME & MAST	`meme.sdsc.edu/meme/intro.html`	Discover motifs in groups of related DNA or protein sequences
Université Libre de Bruxelles, Belgium	`rsat.ulb.ac.be/rsat/`	Analyze regulatory sequences

فصل ششم

کار با توالی پروتئینی

در این فصل می خوانیم:

- ✓ آنچه باید در مورد دامین ها بدانیم (دامین ها: بخشی از توالی و ساختار پروتئین که قادر به انجام یک عملکرد است و بطور مستقل در ادامه ی توالی پروتئین وجود دارد).
- ✓ HMMS، پروفایلها و مجموعه ای از Domain Pfam
- ✓ مشاهده ی (معرفی) سه سایت مشهور برای جستجوی دامینها در پروتئین
- ✓ پیش بینی خصوصیات و ویژگی های فیزیکی ساده در نواحی خاص
- ✓ پیش بینی و شناسائی الگوهای عملیاتی آنزیم پروتئاز
- ✓ شناسائی دامین های قد مانند حاصل از تقابل یک پروتئین دیگر (coiled-coil)
- ✓ شناسائی تغییرات پس از ترجمه

.....

به هنگام مطالعه بر روی یک پروتئین، در حقیقت شما جاپای یک محقق گذاشته اید. بر همین اساس شما در جستجوی یافتن همه چیز هایی هستید که قبل از طراحی آزمایش خود در آزمایشگاه تیار خواهید داشت.

برای مثال، برای اطمینان از اینکه پروتئین موجود در لوله آزمایش دقیقاً همان چیزی است که شما فکر می کنید، باید یک آنالیز فیزیکی ساده بر روی آن انجام دهید. ابزارهای بسیاری بصورت آنلاین وجود دارند که به شما در پیش بینی نتایج آنالیزهای انجام گرفته کمک می کنند و شما را در قدم گذاشتن در میسر درست مطمئن می سازند. در این بخش، نحوه ی یافتن این قبیل ابزارها را آموزش خواهیم داد. همچنین، چگونگی پیش بینی وزن های مولکولی، اندازه گیری نقاط ایزوالکتریک، شمارش اسیدهای آمینه و یا پیش بینی محصول حاصل از عملیات آنزیم پروتئاز را به شما نشان می دهیم. روشن است که علاوه بر وزن و شارژ، یک پروتئین خصوصیات بسیار دیگری نیز دارد اما این دو خصوصیت در درک صحیح از

ویژگی های اختصاصی پروتئین موردنظرتان و نیز آشکار ساختن آن کمک می کنند. شاید مهمترین سوالی که مطرح است این باشد که پروتئین شما به هنگام فعالیت به چه شکل است. آیا پس از ترجمه نیازی به تغییر دارد یا خیر، آیا این پروتئین حاوی نواحی coiled -coil می باشد یاخیر؟ و یا اینکه آیا اصلاً یک پروتئین بین غشائی است. در این بخش، نحوه بکارگیری ابزارهای آنلاین موجود برای پاسخگوئی به این قبیل پرسشها را آموزش می دهیم.

شاید یکی از مهمترین مسائل موجود ساختار باشد. این مسئله به حدی مهم است که دو بخش از این کتاب به آنالیزهای ساختاری اختصاص یافته است (بخش ۱۱ آنالیزاهای ساختاری پروتئین و بخش ۱۲ آنالیزهای ساختاری DNA). به همین دلیل در این بخش کمتر به آن پرداخته شده است. علاوه بر شکل و ساختار پروتئین، تشخیص عملکرد آن نیز مهم است (شناسائی پروتئین و تشخیص عملکرد آن، دو مسئله ی جدا از یکدیگر)، برای مثال: شاید بخواهید بدانیم آیا پروتئین موردنظرتان به کلسیم متصل است و یا اینکه حاوی یک محل فعال آنزیمی است. اگر پروتئین شما یک آنزیم است شکی نیست که باید به سراغ سوبسترای آن بروید(نوعی از مولکول که به جایگاه فعال آنزیم متصل می گردد). برای پاسخگویی به این سوالات، روشی قدرتمند را برای چک نمودن پروتئین به منظور یافتن یک دامین با فعالیت شناخته شده، نشان خواهیم داد. تکنیکهای دیگر برای تشخیص عملکرد یک پروتئین تحت عنوان جستجوهای تشابهات (Similarity searches) آورده شده است. اگردر یک بانک اطلاعاتی یک پروتئین کاملاً شناخته شده با توالی بسیار مشابه با توالی موردنظرتان یافتید آن وقت می توانید بگوئید که آنچه برای آن پروتئین درست است برای پروتئین من هم صدق می کند. اگر ما وارد این قبیل آنالیزهای گیج کننده نمی شویم ولی خیلی هم شما را آسوده نخواهیم گذاشت پس نگران نباشید! ما تمامی آنچه را که برای دانستن جستجوهای مشابه به آن احتیاج پیدا خواهیم نمود را در بخش ۷ این کتاب آورده ایم.

انجام عملیات بیوشیمیائی از طریق کامپیوتر

اگر به انجام عملیات بیوشیمیائی توسط کامپیوتر علاقه دارد، این دو سایت مهیج را به شما پیشنهاد می کنیم:

- ✓ ExPASy (سیستم های تخصصی آنالیز پروتئین[1]) که در آدرس اینترنتی www.expasy.org موجود است و یک صفحه مختص به روش های آنالیز پروتئین دارد.
- ✓ Swiss EMBnet در آدرس اینترنتی www.ch.embnet.org

در حقیقت همان افرادی که بخش بزرگ Swiss-Prot را راه اندازی و برقرار نموده اند، وظیفه ی راه اندازی این دو سایت را نیز بر عهده دارند. این افراد در بررسی عمیق پروتئین ها خبره هستند.

در این قسمت، توجه شما را به پروتئینی جلب می نمائیم که تا کنون آن را ندیده اید و تنها چیزی که از آن می دانید، توالی آن است. شاید این پروتئین محصوص یک پروژه توالی یابی و یا مشابه آن باشد. قبل از باز نمودن و مشاهده درون آن، شاید می خواهید تمامی آنچه که می توانید را در مورد آن بدانید. به همین دلیل است که شروع به مطالعه ای این بخش از کتاب نموده ایم.

اگر پروتئین مورد نظرتان را در اختیار ندارید، اصلاً نگران نباشید. برای هر مثالی که در این بخش استفاده می نمائیم، توالی پروتئینی آن را از Swiss-Prot در اختیارتان می گذاریم . در این صورت می توانید نتایج حاصل از برنامه بکار رفته را با تفاسیر موجود در Swiss-Prot مقایسه نمائید.

[1] *Ex*pert *P*rotein *A*nalysis *Sy*stem

بطور کلی به هنگام استفاده از یک برنامه برای اولین بار، پیشنهاد می شود آن را با یک پروتئین شناخته شده راه اندازی کنید تا از این طریق کارآئی برنامه موردنظر را بررسی نمائید. برای یافتن مثالی مناسب، به سراغ سرور SRS در بخش ۲ بروید. این سرور نحوه بررسی کارآمد بودن یک برنامه را نشان می دهد.

شناسائی و پیش بینی خصوصیات فیزیکوشیمیائی یک پروتئین

Protparam برنامه ای است که از طریق سرور ExPASy بصورت آنلاین اجرا می گردد و در حقیقت روشی مناسب برای شناسائی بسیاری از خصوصیات فیزیکوشیمیائی ساده بدست آمده از توالی پروتئین مورد نظر(که همان پارامترهای مربوط به هر پروتئین است) می باشد. این برنامه کار را سخت و پیچیده نمی کند و کارکردن با آن بسیار راحت است. روش کار آن به این صورت است:

۱. به آدرس اینترنتی: www.expasy.org/tools/#primary بروید.

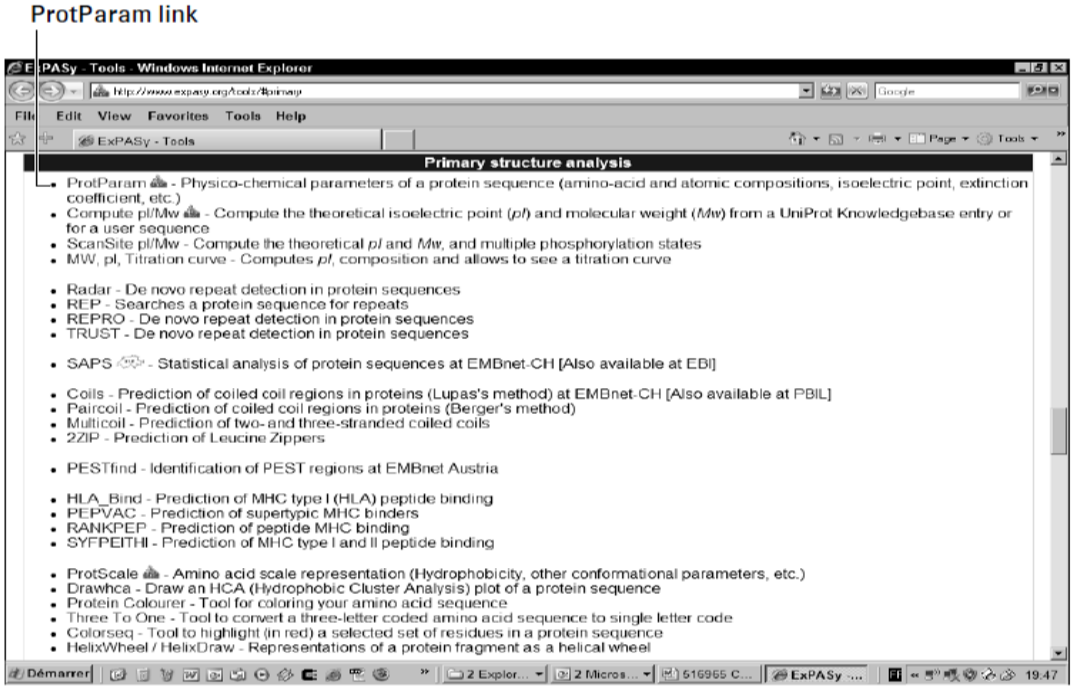
قسمت مربوط به آنالیز ساختار اولیه (Primary Structure Analysis) از مجموعه ابزارهای پروتئومیک ExPASy نمایان می گردد (در شکل ۱-۶ کاملاً مشخص است).

در انتهای این بخش، لیستی جامع از تمامی سایت های در دسترس برای رسیدن به ExPASy، برای مواردی که سایت اصلی در دسترس نیست، در اختیارتان قرار می گیرد.

۲. بر روی لینک ProtParam در بالای صفحه کلیک نمائید. صفحه ی Prot Param Tool نمایان می گردد.

۳. توالی مورد نظر را در بخش جستجو وارد نمائید. برای دستیابی به توالی مورد نظر دو راه وجود دارد:

✓ وارد نمودن شماره ی دسترس مطابق شکل ۲-۶. این روش زمانی ممکن است که توالی شما در بانک اطلاعاتی توالی در Swiss-Prot موجود باشد. شماره دسترس Swiss-Prot معمولا با یک حرف شروع شده که ادامه ی آن ۵ رقم قرار می گیرد. برای مثال در اینجا ما از پروتئین syntanin 1A موش خرمائی (با کد شناسایی P32851) استفاده می نمائیم.

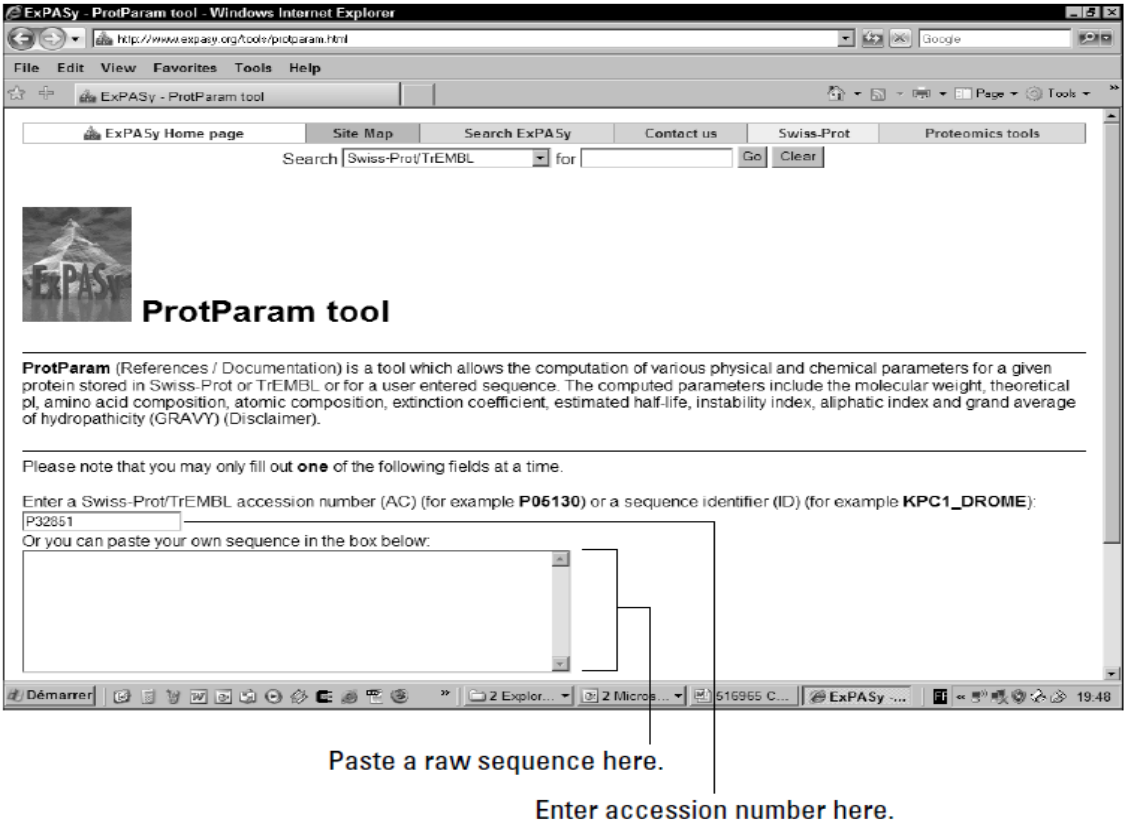


شکل ۱-۶: کار با آنالیز ساختاری اولیه

- ✓ توالی را در فرمت خام خود به پنجره موردنظر انتقال دهید (بار دیگر به شکل ۲-۶ نگاهی بیاندازید). اگر توالی شما از دسته ی توالی های Swiss-Prot نمی باشد و یا اینکه نام توالی خود را نمی دانید، می توانید توالی را مطابق روش بالا منتقل نمائید. توالی موردنظر حتماً باید در فرمت خام خود وارد شود و باید شامل اختصارات استانداردی برای نمایش آمینواسیدها باشد (فاصله ی خالی و شماره ها اگر چه چندان اهمیتی ندارند ولی بصورت اتوماتیک حذف خواهد شد). در صورتیکه توالی شما تحت فرمت FASTA می باشد فاقد خط اول که با یک علامت بزرگتر (>) شروع می شود، می باشد.

۴. بر روی گزینه Compute Parameters کلیک نمائید. در صورتیکه در بخش ۲ توالی را بر اساس شماره دسترسی وارد نموده اید، صفحه بسرعت ظاهر خواهد شد(مطابق شکل ۳-۶).

۵. اگر شماره ی دسترسی Swiss-Prot را وارد نموده و صفحه مطاب شکل ۳-۶ ظاهر شد، آن وقت، محدوده آنالیز را در فیلد های و C-Terminal وارد نمائید (فیلد N-Terminal، در سمت چپ توای پروتئین قرار دارد ، حال آنکه فیلد C-Terminal در سمت راست قرار گرفته است).



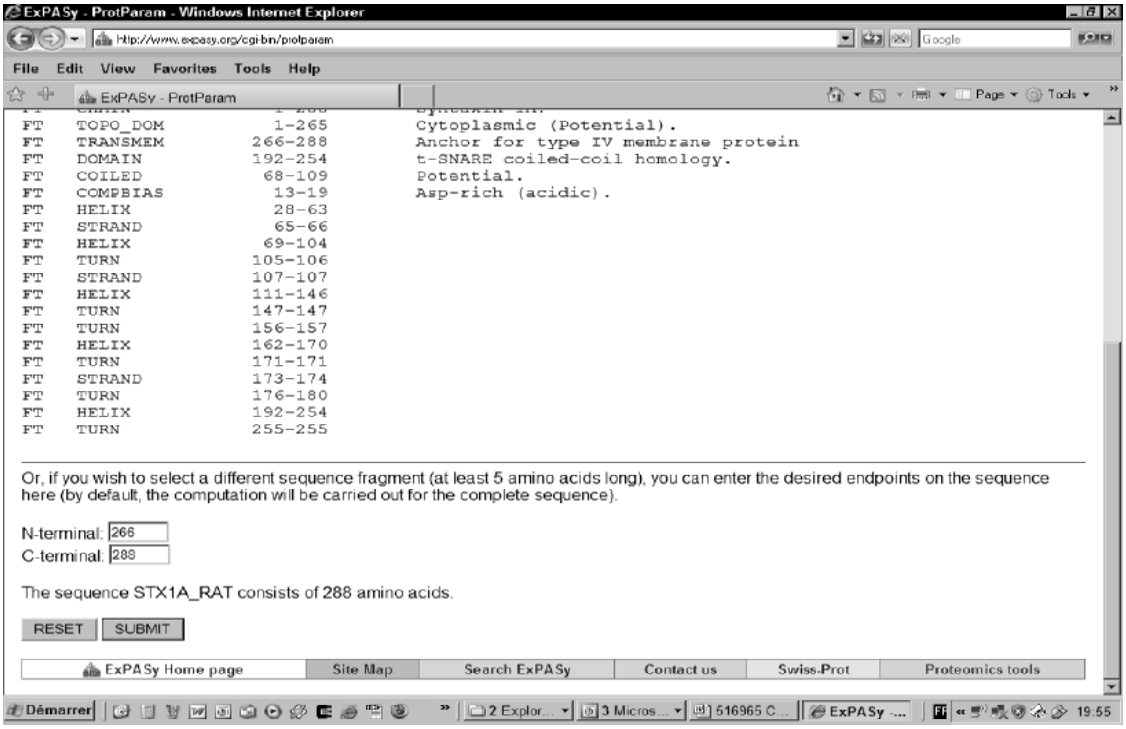
شکل ۲-۶: وارد نمودن شماره دسترسی به درون Prot Param

در بالای صفحه مطابق شکل ۳-۶ ، می توانید لیستی از مشخصات توالی مورد نظر را بیابید. این یک پیش بینی نیست بلکه اطلاعات بدست آمده از ورودی Swiss-Prot است. در قسمت پائین دو پنجره وجود دارد (C-Termianl و N-Terminal) که می توانید آن را با مختصات هر قطعه ای که تمایل دارید، پر نمائید. اگر این دو قسمت را خالی بگذارید آنالیزهای ProtParam بر روی کل توالی صورت می گیرد. در صورت استفاده از مختصات ارائه

شده مطابق شکل ۳-۶، نرم افزار ProtParam قطعه ای از پروتئین را که از آمینو اسید ۲۶۶ شروع شـده و بـه ۲۸۸ ختم می گردد، آنالیز می نماید.

۶. بر روی گزینه Submit به منظور انجام آنالیز کلیک نمائید نتایج حاصل شامل وزن مولکولی، ضریب جذب و نیمـه عمر تخمین بر روی صفحه نمایش ظاهر می شود.

۷. از گزینه ی Save As ← File برای ذخیره سازی و نتایج استفاده نمائید.



شکل ۳-۶: استفاده از محدوده های مختلف در برنامه Prot Param

تفسیر نتایج حاصل از Prot Param

در صورت استفاده از Syntanin به عنوان پروتئین نمونه در اجرای آزمایشی Prot param، مـی توانیـد بـه اطلاعـاتی چـون وزن مولکـولی Syntanin بـصورت تئـوریکی(7:,4 330 دالتـون)، theoretical pI (5.14) و انـدکس آلیفاتیک(۸۳/۹۶) پی ببرید، اینها تنهاگوشه ای از تمامی اطلاعاتی است که این برنامه در اختیارتان می گذارد!

جزئیات بیشتر در مورد چگونگی پردازش این پارامترها توسط برنامه Protparam در آدرس اینترنتی : www.expasy.org/tools/protpar-ref.html موجود است.

وزن مولکولی

برنامه کامپیوتری به راحتی وزن تمامی اسیدهای آمینه را در یک توالی محاسبه می نماید. به یاد داشـته باشـید کـه برنامـه مورد نظر موارد زیر را در نظر نمی گیرد.

- ✓ تغییرات پس از ترجمه نظیر گلیکوزیله شدن و یا فسفریله شدن را در نظر نمیگیرد.
- ✓ این برنامه تغییرات پیچیده ی و تکامل (از قبیل جداسازی پپتید رهبر (lead) و یا برش یک intein) را در نظر نمـی گیرد.

- ✓ برنامه Prot Param حتی نمی تواند تشخیص دهد که آیا پروتئین تکامل یافته بصورت دیمر (اتصال دو پروتئین) ، تریمر، (سه پروتئین) و یادگیر انواع مولتی مر (چند تائی) است.

در حقیقت این قبیل مشکلات می تواند موجب ایجاد تفاوت بارز در نتایج حاصل از آزمایش نسبت به حالت تئوری لذا حواس خود را کاملاً جمع کنید.

ضرایب خاموش سازی (Extinction coefficients)

ضریب خاموش سازی در حقیقت بیانگر مقدار جذب نور توسط پروتئین مربوطه در یک طول موج خاص است (نورمرئی یا غیر مرئی). اندازه گیری این عوامل به هنگام بررسی پروتئین مورد نظر توسط اسپکتروفوتومتر پس از خالص سازی می تواند مفید باشد.

در صورت استفاده از این نتایج، توجه به این نکته حائز اهمیت است که شاخص Prot Param تنها یک نشانه است. برنامه ProtParam به پیش بینی ضریب خاموش سازی از طریق محاسبه ی میزان مشارکت هراسید آمینه به صورت مستقل می پردازد. در محاسبات حاصل، این حقیقت که رفتار اسیدهای آمینه بصورت چشم گیری بر اساس محیط اطرافشان قابل تغییر است، نادیده گرفته می شود. این رفتار قابل پیش بینی نیست و به همین دلیل نتایج حاصل چندان دقیق نیستند. در صورتیکه می خواهید به ضریب خاموش سازی واقعی دست یابید باید آنرا بصورت آزمایشگاهی محاسبه نمائید. به عبارت دیگر، برای اغلب پروتئین ها، ضریب خاموش سازی به ندرت با حالت تئوریکی آن تفاوت فاحشی دارد.

ناپایداری

این پارامتر در حقیقت تخمینی کلی در مورد پایداری پروتئینی مورد نظر در لوله آزمایش است. اگر ضریب حاصل زیر ۴۰ بود پروتئین بطور معمول پایدار است و اگر بالای ۴۰ بود ناپایدار می باشد.

نیمه عمر

پارامتر نیمه عمر در حقیقت یک پیش بینی کلی از زمانی است که نیمی از مقدار پروتئینهای موجود در یک سلول پس از سنتز در سلول ناپدید می گردد. این پیش بینی برای سه نوع ارگانیسم قابل دستیابی است. می توانید براحتی آنرا به ارگانیسم مشابه تغییر دهید.

مقیاس نیمه عمر حاصل در صورتیکه تجزیه پروتئین نتیجه یک پروسه تنظیمی باشد خیلی هم پر معنا نیست. برای مثال، اگر یک سلول بطور اختصاصی یک پروتئاز (پروتئینی که دیگر پروتئین ها را تجزیه می نماید)، را به سمت پروتئین شما سوق دهد، در این صورت پروتئین شما سریعتر از آنکه برنامه ی Prot param به آن اشاره کند از میان می رود.

تجزیه یک پروتئین در کامپیوتر

تجزیه حاصل از پروتئاز (هنگامیکه از یک آنزیم بطور اختصاصی برای ایجاد برش در پروتئین استفاده می شود) می تواند به هنگام مطالعه بر روی قسمت از پروتئین کارساز باشد. همچنین به هنگام انجام عملیات زیر نیز می تواند مفید باشد:

- ✓ جداسازی دامین ها از پروتئین مورد نظر.
- ✓ شناسائی تغییرات بالقوه ی ایجاد شده پس از ترجمه از طریق اسپکتروفتومتری mass
- ✓ جداسازی (حذف) یک پروتئین دنباله (tag) به هنگام ایجاد یک پروتئین اتصالی

- ✓ اطمینان از اینکه پروتئین کلون شده به بعضی از پروتئاز های آندوژن حساس نمی باشد.

خوشبختانه شما می توانید به راحتی مدل های توالی پروتئین را از طریق کامپیوتر بیابید و تمامی اتفاقاتی را که در موقع آزمایش ممکن است رخ دهد، قبل از وقوع مشاهده نمائید. Peptid Cutter ابزاری سودمند برای این منظور می باشد و در سایت تحت وب ExPASy به نشانی www.expasy.org/tools/#proteome قابل دسترس است، حتما سری به آن بزنید، کارکردن با آن بسیار ساده است. توالی خود را Paste کنید و با انتخاب گزینه مورد نظر و کلیک بر روی دکمه، مدلی از پروتئین بریده شده به قطعات کوچکتر و تکه تکه شده با توجه به مشخصات انتخابی در اختیارتان قرار خواهد گرفت.

انجام آنالیز ساختار اولیه (primary structure)

ساختار اولیه در حقیقت توالی آمینو اسیدی پروتئین مورد نظر شما است. به هنگام آنالیز ساختار اولیه در حقیقت تمامی تداخلات بالقوه ی میان اسیدهای آمینه نادیده گرفته شده است. آنالیز ساختار اولیه بر روی یک پروتئین در حقیقت از تداخلات بالقوه ی میان اسیدهای آمینه صحبتی به میان نمی آورد، این دسته از تداخلات را به هنگام پیش بینی ساختارهای دوم و سوم پروتئین مورد نظر پیش بینی می نمایید (فصل ۱۱). آنالیز ساختار اولیه را به سمت شناسایی قطعاتی از پروتئین مورد نظر با نسبت اجزای سازنده خاص سوق دهید. این قطعات در حقیقت خصوصیات جالب توجهی را از پروتئین مورد نظرتان نشان می دهند مانند:

- ✓ نواحی هیدروفیل که در حقیقت قطعات فاصله دار غشا می باشند و خود را به داخل غشا محصور می کنند.
- ✓ نواحی فنر مانند (Coiled-Coil) که نشان دهنده تداخلات و بر هم کنش بالقوه پروتئین با پروتئین است.
- ✓ نواحی هیدروفیلیک که حلقه وار در سطح پروتئین قرار دارند.

روشی که در این قسمت به شما نشان می دهیم اگر چه چندان هم نسبت به روش های هنرمندانه برای شناسائی ساختارهایی دوم (که در فصل ۱۱ به آن می پردازیم) کار آمد نمی باشد ولی هنوز به دلایل زیادی از آن استفاده می شود ؛ مثلا برای سهولت دراستفاده، درک مفاهیم آن و نیز مشاهده ی آنچه به وقوع می پیوندکه این امر از بروز اشتباهات در هنگام کارجلوگیری می نماید.

پنجره های متحرک

تکنیک sinding windows یکی از قدیمی ترین روشهای موجود برای بررسی یک توالی است. اصول کلی آن بسیار ساده است. تنها چیزی که به آن نیاز دارید یک ویژگی شیمیائی و مجموعه ای از مقیاس های مرتبط با هر ۲۰ اسید آمینه است. این ویژگی می تواند شامل هر کدام از پارامترهای فیزیکو شیمیائی قابل اندازه گیری باشد مانند اندازه، قطبیت، هیدروفوبیسیتی و یا حتی گرایش اسیدهای آمینه به سمت حالات ساختاری خاص. مقیاس های موجود در این جدول همه مختص اسید های آمینه می باشند. بسیاری از این جداول بصورت آزمایشگاهی برای هر یک از خصوصیات قابل تعیین است. پس از دسترسی به جدول مورد نظر، کافی است اندازه پنجره را انتخاب و آن را در طول توالی خود گسترده نمائید (slide). هنگامی که پنجره بر روی یک اسید آمینه قرار گرفت، مقیاس های گستره که مرتبط با اسیدهای آمینه است و در محدوده آن قرار دارد، محاسبه و میانگین گیری می شود. مقیاس حاصل مربوط به اسید آمینه مرکزی است و سپس به پنجره اسید آمینه بعدی منتقل می شود. این

پروسه تا انتهای توالی ادامه می یابد. مثالهای آورده شده نشان دهنده چگونگی ایجاد پنجره ها بـر روی یـک تـوالی مـی باشـد. نمای <--X--> نشان دهنده ی اسید آمینه ای است که پنجره بر روی آن متمرکز است.

```
Sequence
    AGVCFGTRESALPTFREDCYGHZPLI
    KJFDESAQZ
      <---A---> Window 1
       <---G---> Window 2
        <---V---> Window 3
```

پس از اتمام عملیات، مقیاس های حاصل از هر اسید آمینه در مقابل توالی مربوطه ترسیم می گردند. زیست شناسان به این نوع از نمایش یک پروفایل مشخصات (Property) می گویند (آن را با پروفایل دامین که در آن فرمولا سیونی از توالی چند تائی هم محور قراردارد اشتباه نکنید)، اگر کمی شانس داشته باشید مـی توانیـد قطعـات بـین غـشائی، حلقـه هـا (سـاختارهای حلقوی) و یا مناطق فنر مانند (Coiled coil) را که با استفاده از این پنجره ها به صورت اسلاید در آمده انـد، شناسـایی کنیـد. هیدروفوبیسیتی یکی از مهمترین آنالیز های موجود است زیراکه نشاندهنده قطعات بین غـشایی و یـا منـاطق مرکـزی در داخـل یک پروتئین است.

بسیاری از روشهای نام برده شده بر اساس تکنیک پنجره های متحرک (sliding windows) می باشند. (به بخش بـالا مراجعه نمائید) پیش بینی های صورت گرفته بر اساس پنجره های اسلاید شده آنقدرها هم حساس و یا دقیق نبوده ولی معتبـر و قابل اعتماد می باشند. در صورت مشاهده ی یک سیگنال قوی به هنگام استفاده از روشی بر ایـن پایـه (Sliding windows) شانس با شما بوده است که توانسته اید به یک سیگنال بیولوژیکی خالص و حقیقی برسید. شاید اصلی ترین فایده ی ایـن روش همین مختصر بودن آن است، به احتمال زیاد تمایلی به تفسیر نتایج برای شما نداشته اند و برای همـین یـک سـیگنال خـام را در مقابل خود می بینید.

چون باید خودتان به تفسیر بپردازید شاید گفتن ۲ قاعده ی ساده که به هنگام تفسیر نتایج حاصل از آنالیزهای بر پایه پنجره های متحرک برقرار است خالی از لطف نیست.

- ✓ بسیار محدود کار کنید و فقط سیگنالهای قوی را شناسائی نمائید.
- ✓ اعتبار توالی خود را چک نمائید. سیگنهای خوب تغییر ناپذیرند و به راحتی با تغییر (کاهش یا افزایش) اندازه پنجـره به اندازه یک اسید آمینه و یا جایگزینی یک جدول مشخصات با جدول مشابه، جابجا نمی شوند.

یکی دیگر از نقاط ضعف این روش اختیاری بودن اندازه پنجره است: چرا ۹ اسید آمینه بجای ۱۳ و یا ۲۱ اسـید آمینـه ؟ هیچ جوابی برای این سؤال موجود نیست اما به هنگام انتخاب اندازه پنجره خود، قانون ساده زیر را به همـراه آن در نظـر داشـته باشید که:

اندازه پنجره باید در همان محدوده ای باشد که اندازه مشخصه ی مورد جستجوی شما در آن قرار دارد.

برای مثال، اگر به دنبال یافتن دامین های بین غشائی هستید که بطور نرمال ۲۱ اسـید آمینـه وسـعت دارنـد، بایـد از انـدازه پنجره ای استفاده نمائید که به این مقیاس نزدیک است. برای مثال ۱۹ یکی از بهترین مقیاس های پنجره برای یافتن نواحی بـین غشائی است.از طرف دیگر، اگر به دنبال مشخصه های ساختاری پروتئین های کروی هستید، رنج بین ۷ تا ۱۱ می تواند مناسـب باشد.

جستجوی قطعات بین غشائی

پیش بینی وجود قطعات بین غشائی می تواند دریافتن چگونگی عملکرد و پیش بینی های ساده دیگر کمک کننده باشد.اگر بدانیم که پروتئین شما بین غشائی است خواهید دانست که نمی توانید بر روی آن با تکنیکهای مشابه پروتئین های گلوبولار کار کنید. در صورت یافتن قطعاتی بین غشائی در بخش انتهائی آمینی یک توالی، می توانید پیش بینی نمائید که پروتئین مورد نظر باید ترشحی باشد. از طرف دیگر، وجود چندین دامین بین غشائی بر روی یک پروتئین نشان دهنده ی یک کانال است. به علت اهمیت این قبیل پیش بینی ها دو راه برای امتحان قطعات بین غشائی نشان می دهیم؛ Protscale (روش ساده تر) و TMHMM (روش کاملتر). TMHMM جزئی از سرور ExPASy نمی باشد، در حقیقت سرویسی است که از طریق کمیته ی ایجاد شده توسط دانشگاه فناوری دانمارک عرضه شده است.

- ✓ Protscale: در حقیقت از تکنیکی مبتنی بر پنجره متحرک و یکی از چندین مقیاس آمینو اسید موجود استفاده می نماید. برای مثال ما از هیدرو فوبیسیتی برای شناسائی قطعات هیدروفونیک که نمایانگر پروتئین های بین غشائی هستند استفاده نمودیم Protscale قادر به پیش بینی همه چیز برای شما نیست، تنها یک پروفایل هیدروفوبیسیتی را به شما نشان می دهد و تفسیر آن را به عهده شما می گذارد.
- ✓ TMHMM در حقیقت برنامه ای هنرمندانه برای پیش بینی قطعات بین غشائی در پروتئین مورد نظرتان است. این برنامه همچنین قادر دادن اطلاعاتی در مورد بخشی از پروتئین که احتمالاً در داخل سلول قرار گرفته و یک بخش از پروتئین که در خارج از سلول قرار دارد، می باشد.

اجرای برنامه Protscale

بیایید شروع کنیم:

۱. به آدرس اینترنتی www.expasy.org/cgi-bin/protscale.pl بروید. صفحه ی اصلی Protscale ظاهر می گردد.

۲. شماره ی دسترسی توالی مورد نظر را در بخش جستجو وارد نمائید. می توانید توالی را در فرمت خام خود وارد نمائید و یا از شماره ی دسترسی Swiss-Prot استفاده نمائید. در این جا شماره P78588، شماره ی دسترسی پروتئین FREL_CANAL (پروتئینی با ۷ قطعه ی بین غشائی) می باشد.

۳. در قسمت پائین صفحه بر روی دکمه رادیویی Hphob./kyte & Doolittle مطابق شکل ۴-۶ کلیک نمائید. برنامه ی Protscale رنجی نسبتاً زیادی از خصوصیات مورد استفاده برای آزمایش بر روی پروتئین مربوطه را در اختیارتان می گذارد. مورد انتخابی توسط ما در این بخش مناسب برای شناسائی زنجیره های بین غشائ است.

۴. در بخش اندازه پنجره، عدد ۱۹ را انتخاب نمائید.

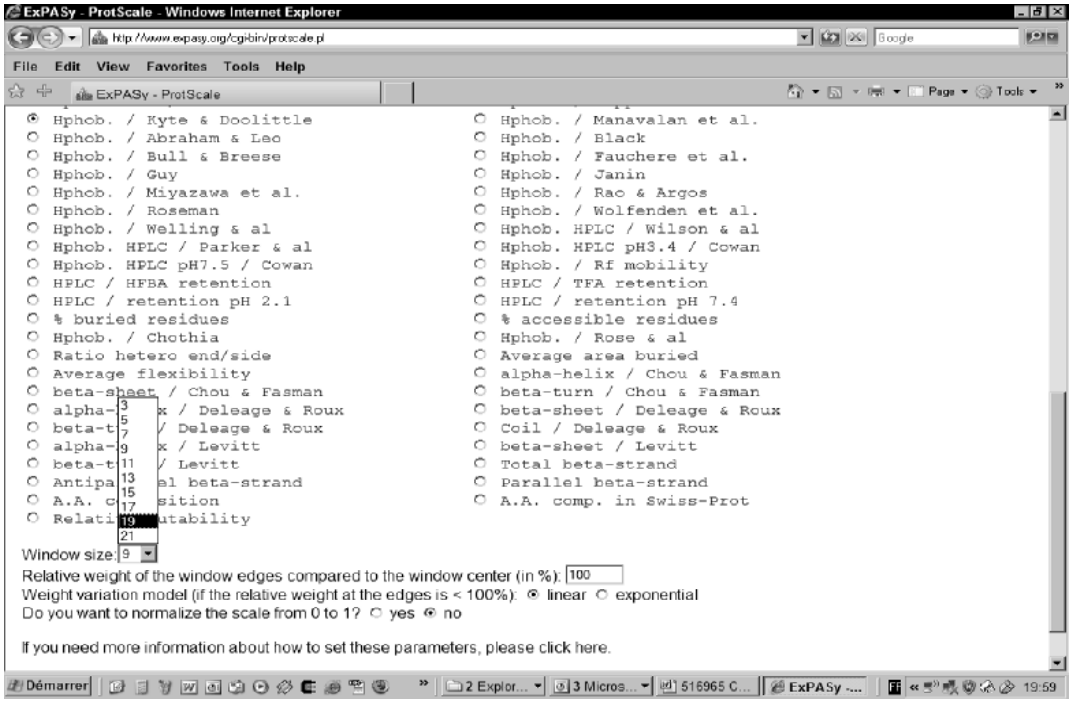
این مقیاس برای جستجوی دامین های بین غشائی مناسب است.

۵. بر روی گزینه Submit در بخش انتهائی صفحه کلیک نمائید.

۶. در صورت استفاده از شماره ی دسترسی Swiss-Prot رنج آنالیز مربوطه را نیز مشخص نمائید.

در صورتیکه درمرحله ۲، شماره ی دسترسی توالی خود را وارد نموده اید، صفحه بسرعت ظاهر می شود (مشابه با شکل ۳-۶).

در بالای صفحه (مطابق با شکل ۳-۶) می توانید لیستی از مشخصه های توالی خود را جستجو نمائید (در حقیقت این عمل یک پیش بینی نیست اما در کنار آن می توان به اطلاعات موجود در ورودی Swiss-Prot دست یافت). در قسمت پائینی دو بخش وجود دارد که در صورت تمایل می توانید آن را با مختصات قطعه ای که بدنبال آن هستید، پر نمائید. در صورت رها نمودن این دو بخش آنالیز های Protparam بر روی کل توالی انجام می گیرد.



شکل ۴-۶: انتخاب یک مشخصه از سرور Protscale

۷. بر روی گزینه Submit برای انجام آنالیز کلیک نمائید. اندکی سعی کنید تا نتایج ظاهر شوند.

۸. به هنگام مشاهده نتایج حاصل، بر روی شکل فرمت GIF در پائین صفحه کلیک نمائید. صفحه ی ظاهر شده به عنوان نتیجه تنها حاوی یک گراف است. از سه فرمت پیشنهادی، GIF
(فرمت قابل تعویض گرافیکی) برای ارائه گراف حاصل مناسب هستند.

۹. برای ذخیره سازی فایل مذکور بر روی Save as ⟵ File کلیک نمائید.

تفسیر نتایج حاصل از Protscale

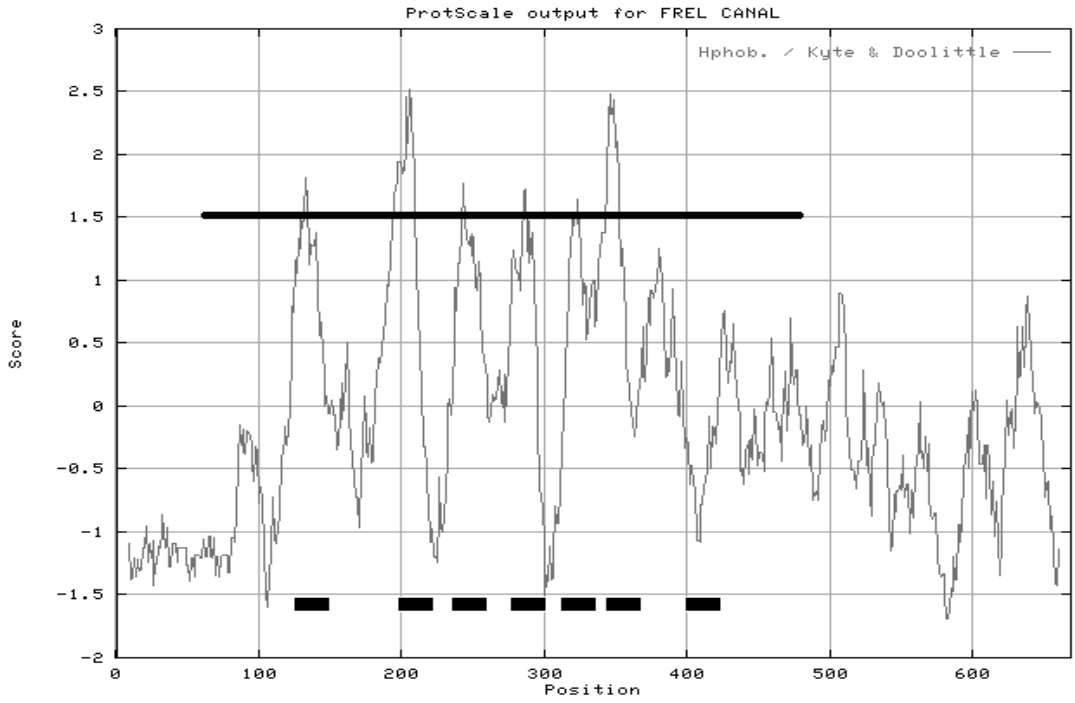
سیگنال مطلوب پایه باید قوی باشد و نسبت به پارامترها (متغیرهای) موجود خیلی حساس نباشد. برای مثال، در شکل ۵-۶، قوی ترین نتایج بدست آمده از پروتئین Swiss-Prot با کد P78588 به هنگام استفاده از مقیاس Kyte and Doolittle Hydrophobicity می باشد. به منظور بررسی میزان کار آمدی نتایج حاصل، مقایسه ای دوباره را این بار با دیگر مقیاس های هیدروفوبیسیتی انجام دهید. مشابه با گستره ی Eisenberg که نتیجه حاصل در شکل ۶-۶ نمایان است. همانطور که مشخص است دو گراف تفاوت اندکی در جزئیات بایکدیگردارند و بوضوح می توان دید که ویژگی های اصلی به خوبی حفظ شده است و پیک های قوی در موقعیت یکسان قرار دارند.

در شکل ۵-۶ برای کمک به تفسیر نتایج حاصل، از یک خط پررنگ استفاده شده است.

مقیاس حد آستانه ی پیشنهادی به هنگام استفاده از Kyte and Dootittle ، ۱/۶ می باشد اما می توانید این مقیاس رویائی را فراموش کنید و از دستور العمل ساده زیر پیروی نمائید:

۱. قطعه ای از کاغذ را بر روی نتایج بدست آمده قرار دهید.
۲. کاغذ را تا حدی پائین آورید که راس پیک های قوی ظاهر شود.
۳. حد آستانه را تا حدی پائین بیاورید که بتوانید پیک های تیز را ببینید.

با استفاده از این روش ساده، می توانید ۵ تا ۷ ناحیه بین غشائی را به وضوح شناسائی نمائید. ششمین ناحیه را با کمی حدس و گمان (معما چو حل گشت آسان شود) اما ناحیه ی هفتم غیر قابل شناسائی است.



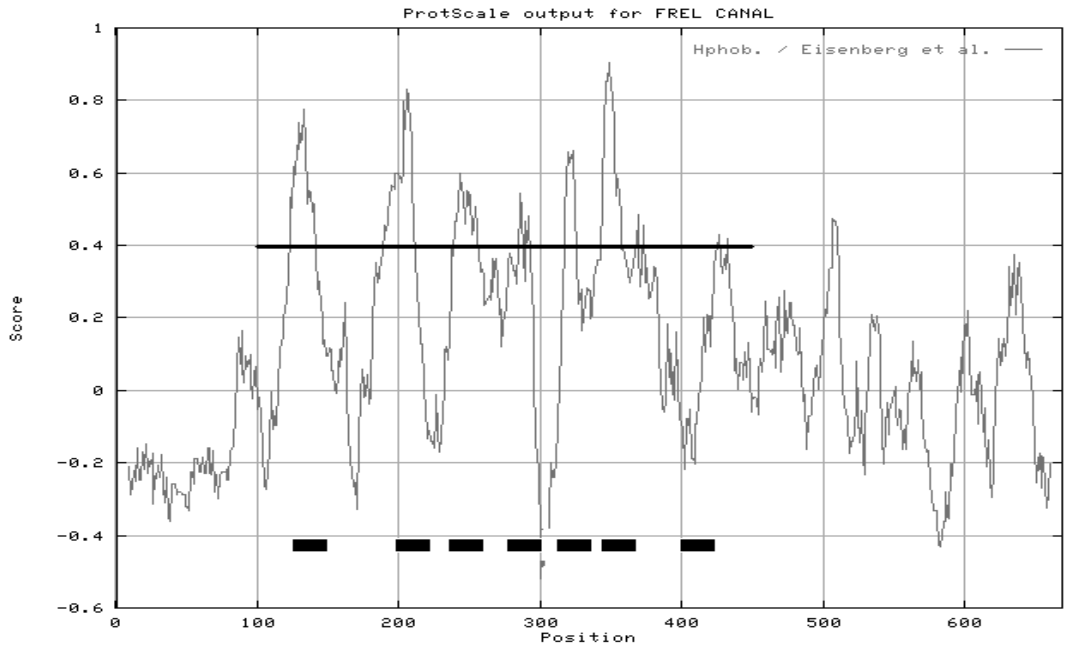
شکل ۵-۶: پروفایل هیدروفوبیسیتی حاصل از Protscale با استفاده از مقیاس Kyte & Doolittle، مستطیل های مشخص شده همان قطعات بین غشائی هستند.

خیلی هم عجیب نیست، تمامی پروتئینهایی که حاوی ۷ ناحیه ی بین غشائی هستند، ۶ تا از آنها به سادگی شناسائی شده ولی هفتمین ناحیه به سخت قابل پیش بینی است.

اجرای برنامه TMHMM

این برنامه از طریق مرکز آنالیز توالی بیولوژیکی (CBS) در دانشگاه فناوری دانمارک قابل اجرا است. CBS (آدرس اینترنتی: www.cbs.dtu.dk/services/) سرویسهای خدماتی جالبی را برای آنالیز توالی در اختیارتان می گذارد. پیشنهاد می کنیم آنها را امتحان نمائید.

TMHMM روشی هنرمندانه که در آن از مدلهای پیچیده و پیشرفته ریاضی (مدلهای Hidden Markov) برای پیش بینی شناسائی نواحی بین غشائی استفاده شده است. آن را با روشی که در ادامه آمده است امتحان نمائید.



شکل ۶-۶: پروفایل هیدروفوبیسیتی حاصل با استفاده از مقیاس Eisenberg

۱. به آدرس اینترنتی http://www.cbs.dtu.dk/services/TMHMM-2.0 بروید. صفحه ی TMHMM از سایت CBS مطابق شکل ۷-۶ ظاهر می گردد.

۲. کمی صفحه را به پائین بیاورید و و توالی مورد نظر را در بخش جستجو TMHMM وارد نمائید. این نرم افزار تنها توالی های با فرمت FASTA را شناسائی می نماید. به عنوان مثال ما از توالی پروتئین FREI-CANAL استفاده نمودیم. برای دستیابی به این توالی از Swiss-Prot عملیات زیر را انجام دهید:

الف) صفحه ی جدیدی را باز نمائید.

ب) به آدرس اینترنتی http://www.expasy.ch/cgi-bin/get-sprot-fasta?FREL_CANAL بروید. این سایت حاوی توالی پروتئین FREL- CANAL با فرمت FASTA می باشد.

ج) توالی مورد نظر را کپی نمائید.

د) توالی را در پنجره ی مربوط به TMHMM منتقل نمائید (شکل ۷-۶).

۳. گزینه های مربوط به تنظیمات خروجی را بدون تغییر رها کنید. (Extensive with Graphics)

۴. بر روی گزینه Submit کلیک نمائید.

۵. پس از ظاهر شدن صفحه ی نتایج، آنها را توسط انتخاب گزینه Save As از منوی File، ذخیره نمائید.

اگر از برنامه ی internet explorer استفاده می نمائید. گزینه Save as را انتخاب نموده و صفحه وب را با فرمت html یاhtm ذخیره نمائید. صفحه ی مربوطه به همراه فهرست راهنما ذخیره خواهد شد. فایل خود را از فهرست راهنما مجزا نکنید.

تفسیر نتایج بدست آمده از TMHMM

اگر نتایج بدست آمده از TMHMM اندکی با نتایج بدست آمده از نرم افزار protscale فرق می کند با اینحال هنوز در یک جهت قرار دارند. این خبر خوبی است زیرا دو روش بکار رفته از اصول متفاوتی پیروی می نمایند. دو فرق اساسی میان این دو نرم افزار وجود دارد:

شکل ۷-۶: صفحه مربوط به TMHMM

۱. TMHMM پیش بینی های دقیق تری انجام می دهد. این پیش بینی ها در قسمت بالائی خروجی قرار دارد درست قبل از ترسیم گراف (نمودار) (plotted graph)(مطابق شکل ۸-۶).

۲. TMHMM قادر به پیش بینی تشخیص قطعاتی است که در داخل سلول و نیز در خارج سلول قرار دارند.

در این مورد، TMHMM قادر به پیش بینی قطعه ی میانی (قطعه ای مابین اسید آمینه ۲۳۴ و ۲۵۵) نبود اما، نتایج خوبی از ۵ قطعه را نشان می دهد. برای هر اسید آمینه از توالی مورد نظر، مطابق نمودار نمایش داده شده در شکل ۸-۶ احتمال قرار گیری آن در ناحیه ی بین غشائی داخل سلول و یا خارج سلول محاسبه می شود.

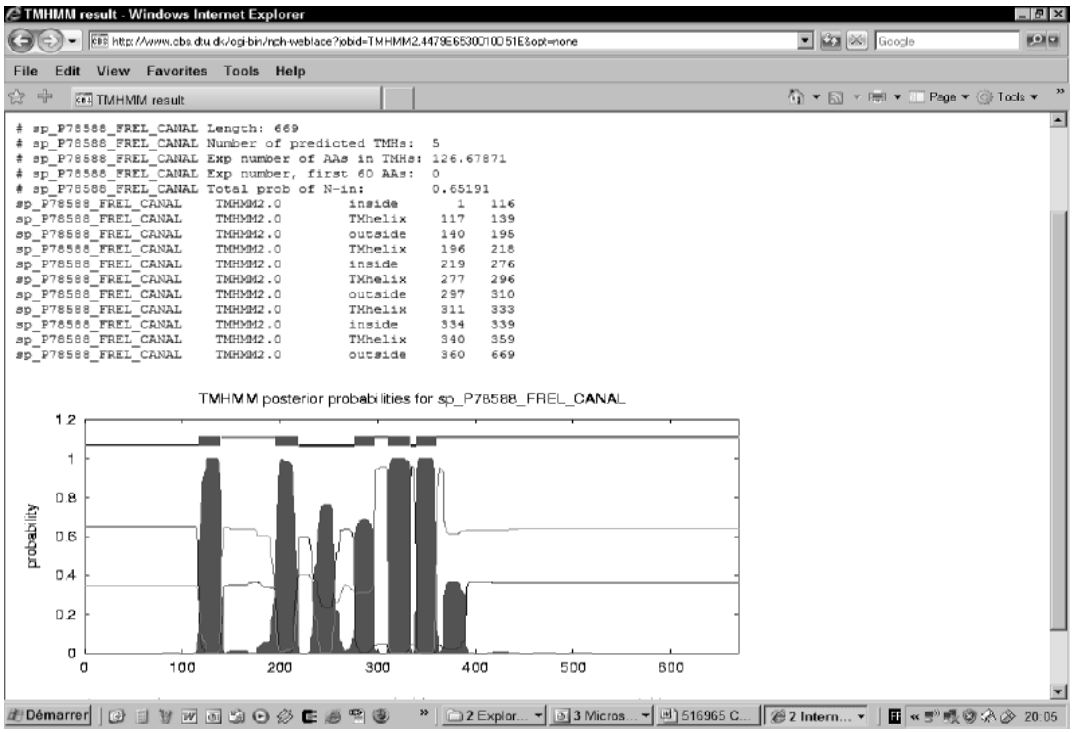
احتمالات نشان داده شده صرفاً یک پیش بینی است و لزوماً قطعی نمی باشد.

اگر به دنبال یک پیش بینی صحیح برای طراحی آزمایش خود می باشید، پیشنهاد می کنیم از روشهای مختلف برای پیش بینی استفاده نمائید و به یکی بسنده نکنید. متدهای مختلف با نتایج مشابه می تواند نشان دهنده این مطلب باشد که شما در راه صحیح قدم گذاشته اید.

جستجوی مناطق فنر مانند (Coiled- coil)

مناطق coiled- coil بخشی از یک پروتئین می باشند که از پیچ و تاب دو یا سه زنجیره α تشکیل شده اند. یکی از دلایل جستجوی این نواحی نقش آنها در بر هم کنش پروتئین- پروتئین می باشد. از دیگر علل موجود این است که این نواحی به هنگام جستجوی یک بانک اطلاعاتی ممکن است منجر به نتایج اشتباه گردند (به فصل ۶ مراجعه نمائید) و شاید به همین دلیل باشد که آن را فیلتر نمائیم. اگر می خواهید به پیش بینی این مناطق در پروتئین مورد نظرتان بپردازید باید ناچار به سراغ سرور COILS در EMBnet بروید:

www.ch.embnet.org/software/COILS_form.html



شکل ۸-۶: نتایج بدست آمده از آنالیز توالی Swiss-Prot FREL- CANAL توسط TMHMM

پیش بینی تغییرات ایجاد شده بر روی یک پروتئین

پس از ترجمه اغلب پروتئین ها پس از ترجمه و قبل از فعالیت در سلول نیاز به تغییر دارند. زیست شناسان به این قبیل اقدامات، تغییرات پس از ترجمه می گویند، شاید دلیل نامگذاری آنها ایجاد تغییرات در زمانی پس از ترجمه (سنتز) پروتئین ها باشد. از جمله ی این تغییرات می توان به اضافه نمودن قندها، تغییرات اعمال شده بر روی اسیدهای آمینه و یا جداسازی و حذف بخش هایی از پروتئین تازه سنتز شده، اشاره کرد. اگر مشغول مطالعه بر روی یک پروتئین جدید هستید، ممکن است بخواهید از این تغییرات آگاه شوید. اهمیت این موضوع زمانی مشخص خواهد شد که بخواهید یک پروتئین انسانی را در باکتری کلون نمائید که علت آن نیز مشخص است، زیرا برای فعال سازی پروتئین مورد نظر، نیاز به تغییرات پس از ترجمه ای است و باکتری قادر به انجام آنها نیست.

یکی از سودمند ترین ابزارهای آنالیز تغییرات پس از ترجمه، PROSITE است (یک بانک اطلاعاتی در سایت ExPASy) این پایگاه شامل لیستی از توالی های الگوی کوتاه (به آن pattern نیز می گویند) می باشد که آزمایشات اغلب با خصوصیات زیستی خاص موجود در آنها مرتبط هستند. از طرفی بسیاری از قبیل نمونه ها (Pattern) با تغییرات پس از ترجمه در ارتباط می باشند. یکی از سرور های ExPASy (www.expasy.org) که از طریق آن می توانید به مقایسه ی توالی پروتئین مورد نظرتان با مجموعه ی الگوهای موجود در PROSITE بپردازید و تغییرات اعمال شده بر روی پروتئین را با آنچه پیش از این رخ داده جستجو نمائید.

به هنگام آنالیز توالی همواره به خاطر داشته باشید: توالی های کوتاه مشابه (مانند آنهائی که کمتر از ۲۰ اسید آمینه دارند) لزوماً دارای عملکرد یکسانی نخواهند بود. بنابراین اگر یک توالی کوتاه عملکردی مانند اتصال دهنده ی ATP را بر عهده دارد و پروتئین مورد نظر که در حال آنالیز آن هستید شامل همان توالی کوتاه است، این بدان معنی نیست که پروتئین شما نیز یک

پروتئین اتصال دهنده ی ATP است. تنها چیزی که مشخص است این است که شما نشانه هائی از امکان وجود یک پروتئین اتصال دهنده ATP را در دست دارد. البته، هر چه قطعه ی مربوطه بلند تر باشد، این نشانه ها نیز پررنگ تر خواهند شد.

از طرف دیگر همین مسئله (احتیاط در مورد مفاهیم مناطق مشابه کوتاه) در مورد الگوهای Prosite نیز صدق می نماید زیرا الگوهای لیست شده اغلب کوتاه می باشند. در ادامه، نحوه ی جستجوی توالی مورد نظر را برای یافتن الگو های شناخته شده ی PROSITE به شما نشان می دهیم.

جستجوی الگوهای PROSITE

ScanProsite در حقیقت یک سرور با قابلیت مقایسه ی پروتئین مورد نظر با لیستی از الگوهای موجود در بانک اطلاعاتی PROSITE می باشد. الگوهای موجود در این بانک اطلاعاتی هر یک به دست متخصصانی آموزش دیده طراحی گردیده است. درصورت یافتن یک الگوی Prosite در پروتئین مورد نظر، در اغلب اوقات می توانید به نشانه ای واضح از عملکرد آن دست یابید.

الگوهای Prosite

هنگامیکه برای اولین بار زیست شناسان اقدام به دستیابی به توالی های پروتئینی کردند، یکی از اولین کشفیات آنها این بود که بعضی از توالی های فشرده کوچک با بعضی از خصوصیات مهم (از قبیل مکان یابی سلولی، اتصال دهنده ی لیگاند و نظایر آن) در ارتباطند.

Amos Bairoch (سازنده Swiss-Prot) که شروع به گسترش این کشفیات و ساخت بانک اطلاعاتی پروتئین خود نمود، برای کمک به سازمان دهی و شناسائی پروتئین ها، مجموعه ای از قطعات فشرده شده و کوچک را برای دسته بندی و آنالیز پروتئین های جدیدتر جمع آوری کرد. بسیاری از این قطعات خاص بنام Pattern (الگو) شناخته شده و امروزه از آنها در سطح وسیعی برای شناسائی و تشخیص پروتئین های جدیدتر استفاده می گردد. PROSITE ، نامی است که Amos بر روی بانک اطلاعاتی الگوهای خاص گذاشت. امروزه پایگاه Prosite علاوه بر الگو های مختلف، حاوی مدل شناخته شده ی جدیدی به نام profile (پروفایل) است. پروفایل ها شامل اطلاعاتی از مکانهای مختلف یک خانواده پروتئینی است و فقط تعدادی از مکانهای فشرده شده ی خاص مطابق آنچه که بخش الگوها عرضه می کند را شامل نمی شود (ضمنا پروفایل های دامین با پروفایل مشخصات که در همین بخش به آن می پردازیم متفاوت است).

اکنون به طریقه ی استفاده از Scan prosite می پردازیم:

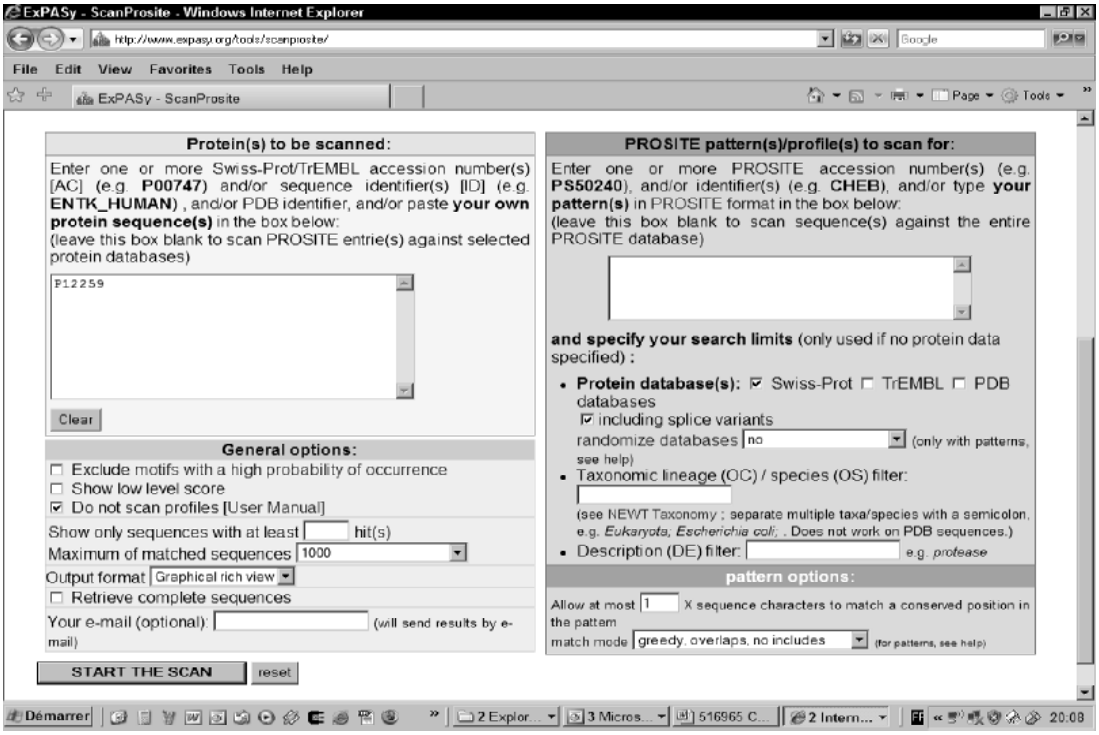
۱. به آدرس اینترنتی www.expasy.org/tools/scanprosite/. بروید. صفحه ی مربوطه از سایت ExPASy ظاهر می گردد.

در صورتیکه صفحه را کمی پائین آورید، به دو بخش کاملاً مشخص برخورد خواهید نمود:

- ✓ بخش سمت راست صفحه (مطابق شکل ۹-۶) زمانی کاربرد دارد که شما یک الگو (pattern) را در اختیار دارید و می خواهید ببینید که آیا سایر پروتئین های موجود در Swiss-Prot این الگو را شامل می شوند یا خیر؟
- ✓ بخش سمت چپ صفحه (مطابق شکل ۹-۶) مربوط به زمانی است که یک توالی پروتئین را در اختیار دارید و می خواهید بدانید کدام یک از الگوهای Prosite را شامل می شود.

۲. توالی مورد نظر و یا شماره دسترسی مربوطه را در قسمت جستجو در سمت چپ صفحه در بخش Protein[s] to be Scanned (پروتئین هایی که قرار است اسکن شوند) وارد نمائید.

- ✓ در این حالت می خواهیم از یک توالی پروتئین به یک الگو برسیم، پس به سراغ بخش سمت چپ صفحه می رویم. برای جستجو به دو صورت می توانید اقدام نمائید.
- ✓ استفاده از فرمت اولیه (اسیدهای آمینه، فاصله ها و اعداد) اشکالی را در سیستم ایجاد نمی کنند و به طور اتوماتیک حذف می گردند.
- ✓ شماره های دسترسی، برای مثال ما از پروتئین Swiss-Prot با شماره های P12259 استفاده نمودیم. این پروتئین در حقیقت محرک تولید فاکتور انعقادی V انسانی می باشد.



شکل ۹-۶: صفحه ی Scan prosite از سرور ExPASy

۳. بخش Exclude motifs with a hign probability of accurrence را بدون علامت رها نمائید. در صورتیکه به تنظیمات دست نزنید، سرور به طور اتوماتیک الگوهای کوچک را که احتمالاً بدون معنی می باشند، نادیده می گیرد. اگر می خواهید این الگوها در نظر گرفته نشوند، این بخش را از حالت انتخاب خارج نمایید.

۴. بخش Do Not Scan Profiles را فعال نمائید. برای رسیدن به اهداف این مرحله از کار، بهتر است از اسکن پروفایلها به دلیل وقت گیر بودن خودداری کنید. به علاوه، در هنگام برخورد با تغییرات پس از ترجمه الگوها نسبت به پروفایلها ارجحیت بیشتری دارند.

۵. بر روی گزینه Start the scan کلیک نموده، اگر کمی صبر کنید، صفحه ی نتایج مطابق شکل ۱۰-۶ ظاهر می گردد. منتظر نتایج باشید! بر روی لینک های دیگر تا ظاهر شدن نتایج نهائی کلیک نکنید. سرور Scan prosite گاهی آهسته کار می کند. اگر نتایج مربوطه را هرگز مشاهده نکردید ناچار باید از سایت ExPASy که سرعت بهتری دارد استفاده نمائید. به لیست موجود در انتهای این بخش سری بزنید.

۶. با انتخاب گزینه Save As از منوی File، برای ذخیره سازی اطلاعات بدست آمده اقدام نمائید.

تفسیر نتایج حاصل از Scan Prosite

اصلی ترین مشکل به هنگام مواجهه با اغلب الگوهای پس از ترجمه این است که آنها کوتاه هستند تطابق توالی های کوتاه با الگوهای مربوطه اغلب بر اساس شانس بوده و لذا نمی توان انتظار علائم بیولوژیکی حقیقی را داشت. قانون حاکم بر این مسئله آن است که مشابهت های مابین توالی های کوتاه باید همواره با شک و تردید بررسی گردد. یک راه خوب برای جلوگیری از بسیاری تغییرات غیر مشابه خواندن با دقت مستندات مرتبط با الگوها (PDOC) است.

درک صحیح از خروجی Scan Prosite

اولین بخش از خروجی بدست آمده بنام Hits by pattern معروف بوده و نمایانگر مکان بسیاری از انواع الگوهای بدست آمده یافته شده در پروتئین مرتبط است. در حقیقت خلاصه ای است از اینکه هر کدام از خانواده الگوها که رنگی مخصوص بخود را دارد، در چه محلی قرار گرفته است. نشانگر موس خود را بر روی یکی از مربع های رنگی گذاشته تا نام مربوط نمایان شود.

در بخش Hits by patterns جزئیات مربوط به هر الگوی یافته شده در توالی را مشاهده نمائید. از طریق لینک های موجود در خروجی می توانید مستند سازی کنید. بخشی از این خروجی در شکل ۱۰-۶ قابل مشاهده است. برای هر الگوی موجود در Prosite، برنامه ی Scan Prosite اطلاعاتی مشابه آنچه در ادامه آمده است ارائه می دهد:

- ✓ **Ps # (5)** اولین لینک برای مستند سازی الگوی مربوطه است. این مستند سازی که با کیفیت بسیار خوب انجام می پذیرد (شامل اطلاعات بسیار زیادی از الگوی مربوطه و فعالیت بیولوژیکی بالقوه ی آن است.
- ✓ **نام الگوی مربوط** بصورت پررنگ (Bold) نمایان است.
- ✓ **Structure (ساختارها):** تعدادی از الگوهای یافته شده در پروتئین ها به همراه ساختار سه بعدی شناخته شده ای قابل مشاهده هستند. اگر الگوی شما نیز جزء این دسته باشد می توانید نام PDB ساختاری آن را به همراه لینکی برای نمایش ساختار سه بعدی در این محل پیدا نمائید.

PDB یک بانک اطلاعاتی (که در حقیقت یک بانک اطلاعاتی سه بعدی شناخته شده پروتئینی است) می باشد. برای کسب اطلاعات بیشتر در مورد PDB به فصل ۱۱ مراجعه نمائید.

در صورت کلیک بر روی هر یک از این لینکها، شکلی با فرمت GIF از ساختار مربوطه ظاهر می گردد. شکل ۱۱-۶ با کلیک بر روی لینک 1czv ظاهر می شود.

برای مشاهده می توانید به آدرس اینترنتی زیر مراجعه نمائید.

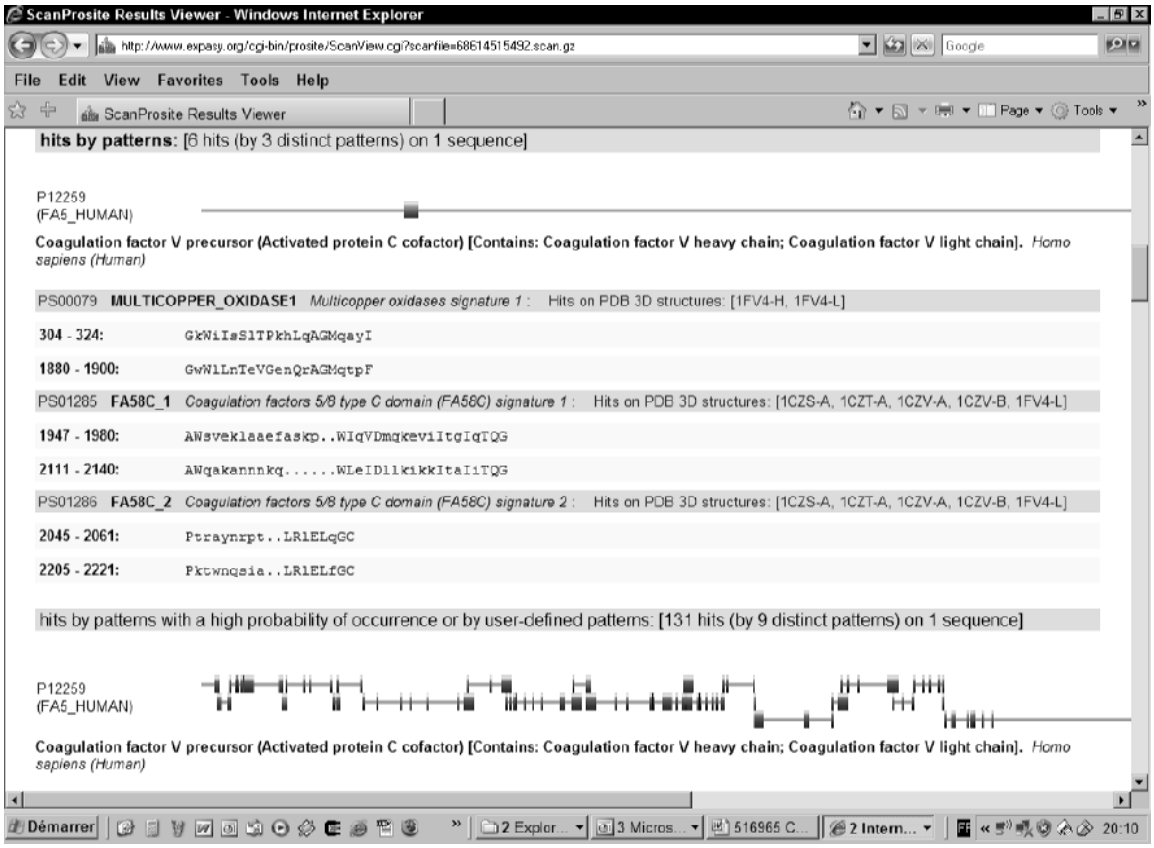
www.expasy.org/ cgi-bin/view-pdb?pdb=1czv&ps=PS01285

- شکل مشاهده شده شامل ساختار سه بعدی از پروتئین در یک نمای ربان شکل است.
- مکان این الگوی خاص (بصورت سبز رنگ) در ساختار مربوطه قابل مشاهده است.

- ✓ **The list**: این لینک لیستی از قطعاتی را که حاوی الگوهای موجود در پروتئین مربوطه است، ارائه می دهد. اعداد نمایانگر مکان تطابق ایجاد شده در توالی مربوطه است. حروف بزرگ نشان دهنده ی اسیدهای آمینه ای است که از

طریق الگوی مربوطه مجزا وتوصیف شده اند. از طرف دیگر حروف کوچک نشان دهنده اسیدهای آمینه ای است که از این قانون مستثنی می باشند.

خروجی بعدی(با عنوان Hits by patterns with a High probulity of occurance) مشابه با همین بخش بوده و نشان دهنده ی مکان الگو های کوتاه و متداول در توالی مورد نظر است.



شکل ۱۰-۶: خروجی برنامه ی Scan Prosite

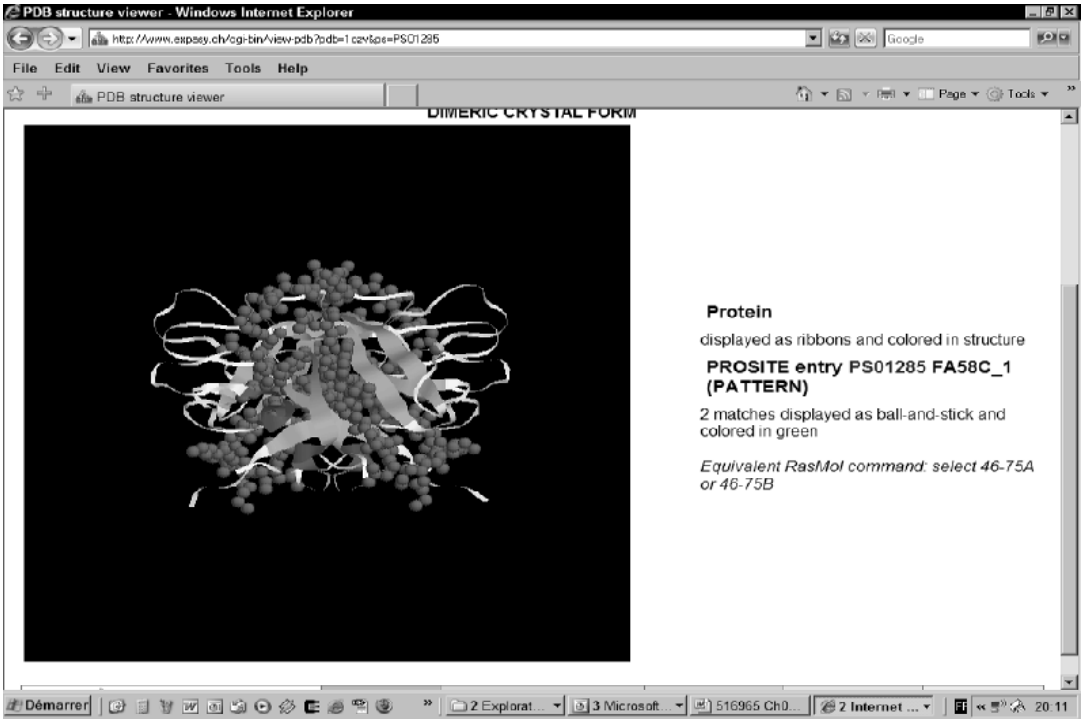
مراقب الگوهای(pattern) کوتاه باشید

خروجی کامل بدست آمده از P12259 شکلی مناسب از نامناسب بودن اکثر الگوهای کوتاه بدست آمده ارائه می دهد. این خروجی به وضوح نشان می دهد که ۱۹ محل از myristillation در فاکتور انعقادی V انسانی وجود دارد. حال به نظر شما می توان به این اطلاعات اعتماد نمود؟ در صورت کلیک بر روی لینک Ps0008 در انتهای بخش (High Hits by Patterns with a Probability of Occurrence) اطلاعاتی ظاهر می شود مبنی بر اینکه Myristate در حقیقت یک اسید چرب است که به بخش انتهای آمینی پروتئین (در انتهای سمت چپ توالی) متصل شده است و موجب اتصال آن به غشا می شود. هیچکدام از اتصالات گزارش شده بر روی انتهای آمینی پروتئین نبوده است بنابراین به وضوح مشخص است که هیچکدام از تطابقات بدست آمده خالص نمی باشند.

استفاده از اطلاعات گونه ها

به هنگام یافتن تغییرات پس از ترجمه، اطمینان پیدا کنید که این تغییرات با گونه هایی که توالی ها از آن ایجاد می شوند، سازگار است. تغییراتی که در پروکاریوتها به وقوع می پیوندند با تغییراتی که در یوکاریوتها ایجاد می شوند متفاوت است، شما

باید این اطلاعات را در PDoc مرتبط با الگوی مربوطه پیدا نمائید. برای مثال، پدیده ی myristillation تنها در یوکاریوتها اتفاق می افتد. درصورت یافتن محلی از وقوع این پدیده در پروتئین پروکاریوتی شما باید آنرا نادیده بگیرید.



شکل ۱۱-۶: مکان یابی الگوی prosite (در شکل سه بعدی ارائه شده از ساخت یک پروتئین

سیگنالهای ضعیف با هم ایجاد سیگنالهای قوی تر می کنند

فراتر از آنچه که می توان از یک خروجی Scan prot انتظار داشت، ارائه لیست الگویی شامل دو انطباق کد مختص فاکتورهای انعقادی می باشد: FA58C-2 , FA58C-1

اگر به آنها بصورت تکی نگاه کنید هیچکدام از آنها خیلی بارز و آشکار نمی باشند. از طرف دیگر این حقیقت که دو الگوی مرتبط با هم را در نزدیکی یکدیگر یافته اید بسیار هیجان انگیز است زیرا نشان دهنده این است که هر دو الگو به احتمال زیاد خالص و حقیقی هستند. این مثال خوبی از این حقیقت است که دو نشانه ی ضعیف، در صورتیکه پایدار و مستقل باشد، می تواند یک نشانه ی قوی را درست نماید. درست مانند زمانی که دو شاهد نزدیک بین از مشاهده یک پدیده سخن بگویند. برای مشاهده گران و محققین این امر نسبت به حالتی که یک شاهد با چشمانی چون عقاب از وقوع یک پدیده سخن بگوید از اهمیت بیشتری برخوردار است. البته به این شرط که دو مشاهده گر نزدیک بین هر گز با هم در ارتباط نبوده باشند و مستقل باشند.

حذف الگوهای ضعیف

به هنگام در اختیار داشتن تعداد زیادی الگوی ضعیف، استراتژی مناسب برای مواجه شدن با آنها استفاده از ساخت یک هم ترازی چند تائی توالی با استفاده از توالی های مرتبط است.

یک الگو در صورت مرتبط بودن با یک تغییر پس از ترجمه حقیقی و خالص می تواند به خوبی ذخیره شود. فصل ۲ را به توضیحات کامل در خصوص هم ترازی های چند تائی اختصاص داده ایم. بعلاوه همه ی آنچه را که باید برای ساخت هم ترازی های چند تائی توالی با کیفیت عالی بدانید در فصل ۹ موجود است.

همه ی مطالب را در Prosite نمی توان یافت

شاید گفتن این نکته از اهمیت بالائی برخوردار باشد که مجموعه الگوهای Prosite حاوی توضیحاتی پیرامون تمامی تغییرات پس از ترجمه معروف نمی باشد. اگر قادر به یافتن مطلب مورد نظرتان در prosite نیستند، سرور ExPASy لینکهای زیادی را برای اتصال به ابزارهای دیگر که مختص این قبیل آنالیز ها هستند در اختیارتان می گذارد (لینکهای مورد نظر در بخش پیش بینی تغییرات پس از ترجمه از سرور prosite و به آدرس www.expasy.org/tools/#ptm قابل دسترس است).

یافتن دامین های معروف و شناخته شده در پروتئین

اساتید دامین را با عنوان " واحدهای پیچ خورده ی کروی مستقل" معرفی می نمایند. شاید بتوان گفت یک دامین بخشی از یک پروتئین می باشد که قادر به حفظ شکل خود به هنگام جدا شدن از پروتئین است. یک دامین حاوی حداقل ۵۰ اسید آمینه است. دامین ها مشابه با وسایل آشپزخانه ی شما هستند: آون، میکروویو، یخچال برقی و نظایر آن تمامی این وسایل با هم یک آشپزخانه را تشکیل می دهند در حالیکه هر کدام بصورت جداگانه قرار گرفته اند. شما می توانید میکروویو را از آشپزخانه ی خود خارج نمائید، البته تنها درصورتیکه مجبور به استفاده از یک وسیله ی ساخت ذرت بوداده در آزمایشگاه خود هستید.

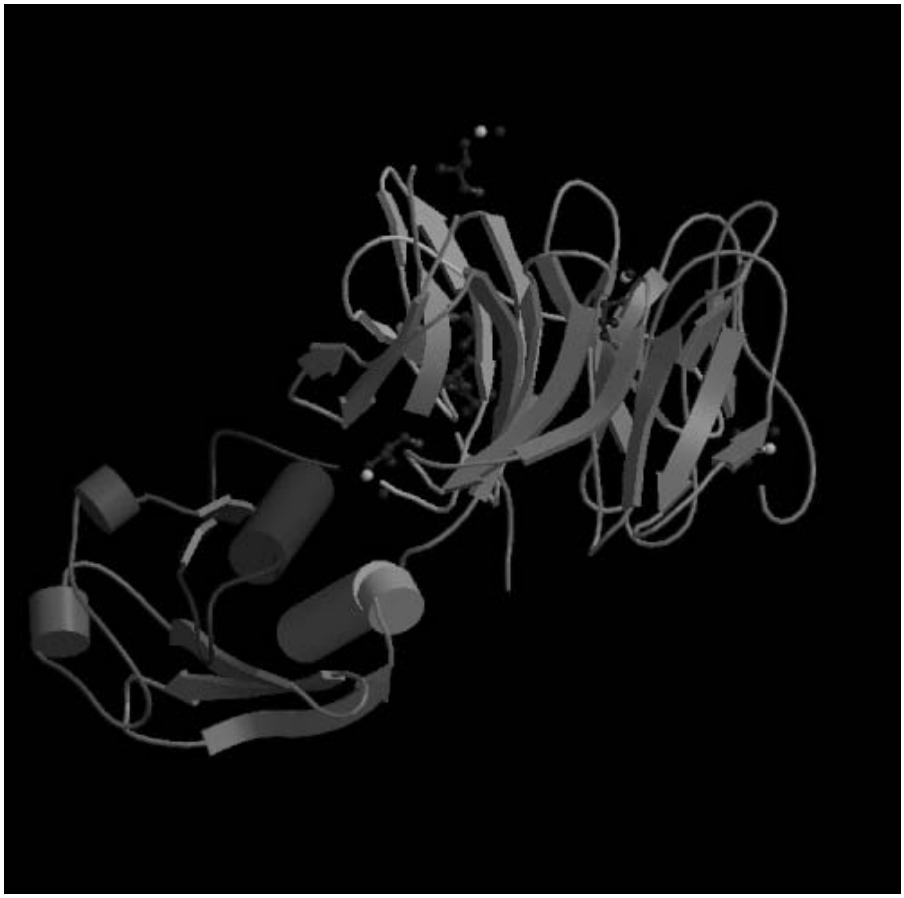
پروتئین مورد نظر شما شامل ۲ یا ۳ دامین می باشد. شکل ۱۲-۶ نمایشی لحظه ای از پروتئین TolB است (پروتئینی با ۲ دامین). بطور معمول هر دامین نقش خاصی را در پروتئین ایفا می کند. از طرفی ممکن است با دیگر پروتئینها نیز در تماس باشد و یا به یک یون مانند کلسیم و یا روی متصل شود و همچنین ممکن است یک مکان فعال را در برگیرد. داشتن یک دامین با فعالیت کاتالیکی که با یک دامین پیوند زننده و یک دامین تنظیم کننده در ارتباط باشد امری متداول است. تصور کنید در جستجوی یک سرخ کن هستید و کباب پز (کاتالیز کننده) و نگهدارنده سرخ کن (پیوند زننده) و کلید (تنظیم کننده) را در اختیار دارید.

هم در عمل و هم در اصول موجود، بین جستجوی الگوها (Pattern) و دامیندر توالی مورد نظر اندکی تفاوت وجود دارد. از طرف دیگر هنگامیکه مشخصات یک دامین بیان شود، آنگاه یافتن این مسئله که پروتئین شما حاوی این دامین هست و یا خیر کار آسانی است. مجموعه های بسیاری از دامین ها در سراسر دنیا وجود دارند (جدول ۱-۶). در بسیاری از این مجموعه ها، گروهی از متخصصان یک به یک دامین ها را هنرمندانه آفریده اند دقیقاً مشابه ساخت ساعتهای سویسی. این مجموعه ها اغلب، نتایج دقیقی را در اختیارتان می گذارند، البته ممکن است در بعضی مواقع نقص های کوچک وجود داشته باشد. (به جدول ۱-۶ نگاهی بیاندازید). کامپیوتر تمامی توالی های شناخته شده را ذخیره و آنها را در دسته هایی از دامین ها طبقه بندی می کند. این قبیل مجموعه ها بسیار گران بها بوده و اما اغلب قابلیت اطلاع رسانی کمتری را دارند (جدول ۱-۶).

انتخاب مجموعه ی صحیحی از دامین ها

دامین ها در بیولوژی از اهمیت خاصی برخوردار بوده و لذا زیست شناسان اغلب تلاش کرده تا مجموعه ی شکیلی ازآنها را در اندک زمان ممکن و مطابق با پیشروی زمان پدید آورند. در این راه با دو مشکل مواجهند: ۱) شناسائی یک دامین خاص بصورت روشن و واضح کار ظریفی است و نیاز به مهارت دارد ۲) اغلب اساتید دریافتن آنها با یکدیگر اتفاق نظر ندارند (اکنون متوجه شدید که چرا محققین و نویسندگان مقالات اکثرا دامین های مهم را با روشهایی متفاوت توضیح می دهند).

از نظر ما، نتایج و تفاصیل موجود همانی است که بیشترین تکرار را در این ۸ مجموعه ی بزرگ از دامین ها دارا است و در جدول ۱-۶ آورده شده است. شاید به ظاهر ساده باشد که بگوئیم مجموعه ی TastyDom بهترین است و دیگر نیازی به استفاده از مجموعه ی POO POO DOM نیست. اما این روش صحیحی نیست. هر کدام از این مجموعه ها معایب و محاسن خاص خود را دارد و بنابراین اگر واقعاً به دنبال درک صحیح از پروتئین خود هستید باید سراغ تمامی این مجموعه ها برویم. البته این کار زمان بر است ولی در نهایت نتیجه آن را خواهید دید.



شکل ۱۲-۶: پروتئین TolB که از دو ناحیه ی دامین ساخته شده است.

پروژه های Interpro امکان جستجوی بسیاری از پایگاههای دامین را در یک زمان با هم فراهم ساخته است. در جدول ۱-۶ (IP) نشان دهنده ی بانک اطلاعاتی دامین است که از طریق Interpro قابل دسترسی است. Interpro کمک می کند تا هیچ داده ای حتی کوچکترین آن را هم از قلم نیاندازید. از طرف دیگر براحتی می توانید از عدم تطابق بعضی از مجموعه های دامین با توالی مورد نظرتان اطمینان حاصل نمائید. متأسفانه Interpro خیلی هم جامع و فراگیر نیست. تنها راه تکمیل آنالیز دامین های موجود در توالی مورد نظر استفاده از سه سرور جامع دامین است:

- ✓ Interproscan
- ✓ CD- several
- ✓ Motif scan

احتمالاً یکی از این سه سرور اطلاعات مورد نظر را ارائه می دهندو این بدان معنا است که شما باید هر سه سرور را یکی پس از دیگری امتحان نمائید. در ادامه ی این بخش نحوه ی استفاده از هر یک سرورها و نیز یافتن بهترین آنها را به شما نشان می دهیم.

Name	Web Address	Number of Domains	Generation
PROSITE-Profile (IP)	www.expasy.org/prosite	616	Manual
PfamA (IP)	www.sanger.ac.uk/Software/Pfam	7973	Manual
PRINTs (IP)	www.bioinf.man.ac.uk/dbbrosers/PRINTS	1900	Manual
PRODOM (IP)	protein.toulouse.inra.fr/prodom/current/html/home.php	736000	Automatic
SMART (IP)	smart.embl-heidelberg.de	685	Manual
COGs	www.ncbi.nlm.nih.gov/COG/new/	4852	Manual
TIGRFAM (IP)	www.tigr.org/TIGRFAMs	2453	Manual
BLOCKs	blocks.fhcrc.org/	12542	Automatic

جدول ۱-۶: نمونه ای از مجموعه های اصلی دامین

یافتن دامین ها با استفاده از برنامه Inter ProScan

اگر قصد استفاده از برنامه Inter ProScan ، برای آنالیز یک توالی پروتئین را دارید و می خواهید ببینید که کدام دامین در توالی مربوطه وجود دارد. با استفاده از برنامه Inter pro scan می توانید توالی مورد نظر را با بانک اطلاعاتی Inter pro scan (یک بانک اطلاعاتی دامین که حاوی مجموعه ای آنلاین از اکثر دامین های موجود است) مقایسه نمائید. این برنامه ابزاری برای یافتن توالی های پروتئین دامین های مربوطه از طریق پایگاه تحت وب است. در ادامه نحوه استفاده از Inter pro scan به طور کامل شرح داده شده است:

۱. به آدرس اینترنتی www.ebi.ac.uk/InterProScan/. بروید. صفحه ی اصلی ظاهر می گردد.

۲. توالی مورد نظر را در بخش جستجو وارد نمائید. در این بخش اغلب فرمتهای قانونی موجود به همراه فرمت اولیه کار قابل شناسائی است، ولی در این بخش از شماره ی دسترسی نمی توان استفاده کرد. در مثالی که در ادامه ی کار آورده شده، سعی شده است از توالی پروتئین FOSB-HUMAN استفاده نمائیم. برای دستیابی به این توالی از Swiss-Prot مطابق دستورالعمل زیر عمل نمائید.

a. یک صفحه اینترنت جدید باز کنید.

b. وارد آدرس اینترنتیwww.expasy.ch/cgi-bin/get-sprot-fasta? FOSB_HUMAN شوید. این سایت حاوی پروتئین FOSB- HUMAN در فرمت FASTA است.

c. توالی مربوطه را کپی نمائید.

d. توالی مورد نظر را به بخش جستجوی inter ProScan منتقل نمائید (مطابق شکل ۱۳-۶)

از طرفی دیگر می توانید یک توالی نوکلئوتیدی را جستجو نمائید. اگرقصد این کار را دارید، DNA مربوطه را بجای PROTEIN (از انتخاب گری که درست بالای توالی است) انتخاب نمائید.

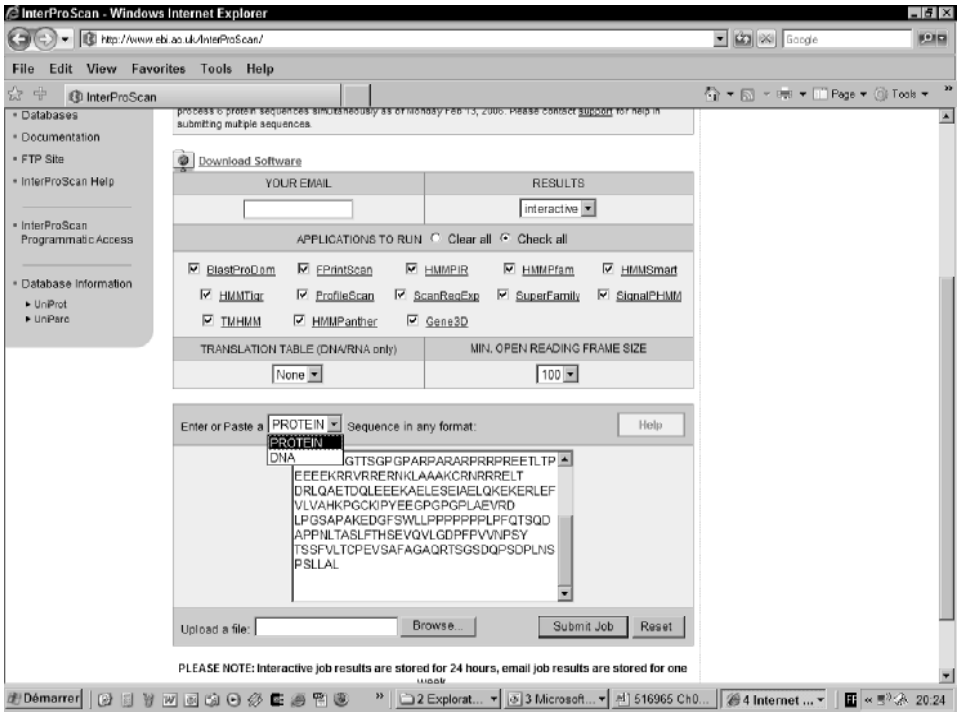
۳. در قسمت Application to Run، بانک اطلاعاتی دامین مورد نظرتان را انتخاب نمائید.

Interpro ، امکان جستجوی بانک های اطلاعاتی دامین را به شما می دهد. از طرف دیگر امکان اجرای برخی از کاربردهای پیش بینی خاص را نیز مانند TMHMM برای پیش بینی های بین غشائی و signal PHMM (برنامه ای برای پیش بینی و تشخیص وجود یک پپتید سیگنال در توالی (قطعه ی کوتاهی در انتهای آمینی پروتئین که موجب می شود تا پروتئین شما به مقصد نهائی خود در سلول برسد دقیقا مانند یک کدپستی) را ارائه می دهد.

می توانید برای سرعت بخشیدن به کار خود برخی از بانک های اطلاعاتی بسیار بزرگ (مثل Pro Dom) را حذف نمائید.

۴. بر روی گزینه Submit کلیک نمائید.

پس از آن لطفاً بر روی هیچ دکمه ای تا دستیابی نتایج کلیک نکنید. بااستفاده از سیستم بازیابی توالی[1] (SRS) نتایج بصورت اتوماتیک نمایان می شود.



شکل ۱۳-۶: صفحه ی اصلی Inter ProScan

۵. نتایج حاصل را ذخیره کنید. برای ذخیره ی نمایش رنگی حاصل بهترین و ساده ترین روش کلیک بر روی دکمه Prt Sc (Print Screen) و سپس کلیک بر روی دکمه Ctrl-V می باشد (با این روش آن را در برنامه مورد

[1] the Sequence Retrieval System

علاقه خود، مثلا Power point، منتقل نمائید). در صورت تمایل به نگهداری نتایج مورد نظر خود در فرمت های دیگر بهتر است آن را به حالت اولیه خود با دو روش زیر ذخیره نمائید:

الف) بر روی گزینه Raw output کلیک نمائید.

صفحه ای جدید ظاهر می شود که اطلاعات را بصورت خام در اختیارتان می گذارد.

ب) از گزینه Save As در منوی File برای ذخیره سازی استفاده نمائید.

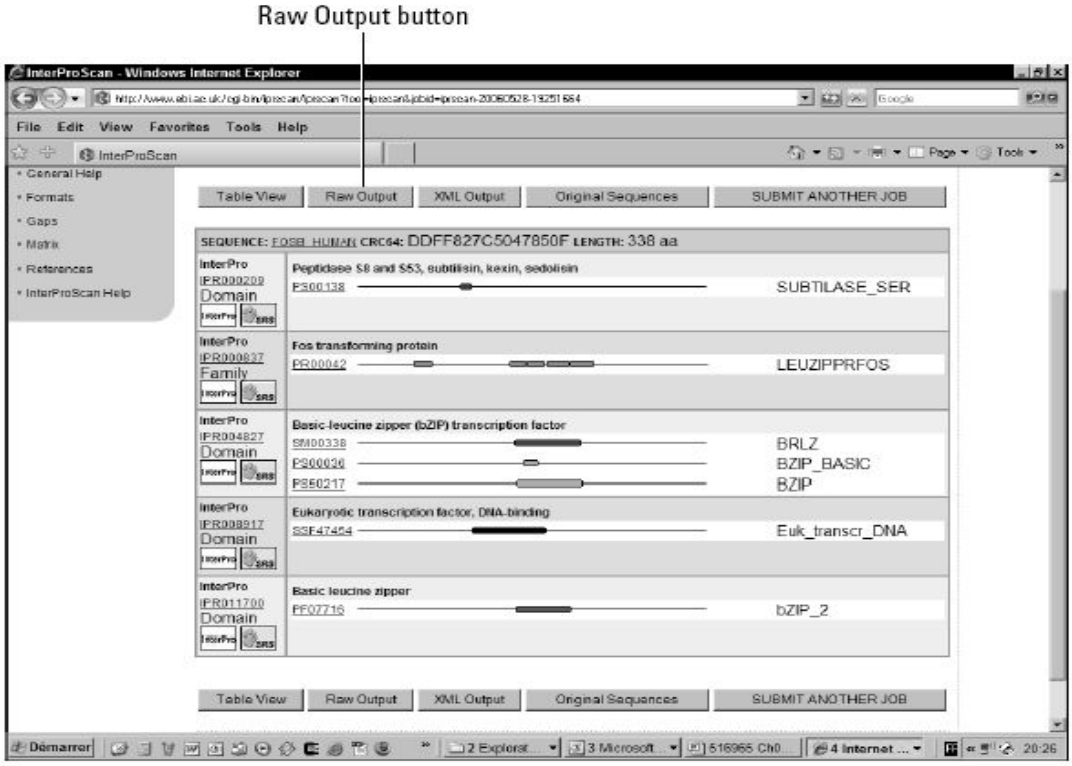
تفسیر نتایج InterProScan

به دلیل گستره ی اطلاعات بدست آمده از Interpro scan تفسیر این نتایج نیاز به مهارت بالائی دارد. قبل از کلیک بر روی هر کدام از لینک ها، مطمئن شوید که از تمامی اجزای صفحه ی نتایج اطلاع دارید.

درک صحیح از جزئیات مربوط به خروجی Inter ProScan

شکل ۱۴-۶ نتایج حاصل از اسکن پروتئین FOSB_HUMAN را از طریق سرور Interpro scan به شما نشان می دهد. این برنامه نشان دهنده ی نتایج گوناگونی است که به هنگام مقایسه یک توالی در بانک اطلاعاتی دامین انتظار دارید. در یک خروجی هر خط نشان دهنده ی یک تطابق با برخی از انواع دامین های Interpro است:

- ✓ اولین ورودی در هر ستون نشان دهنده ی نوع تشخیص داده شده است: خانواده (Family) و یا دامین. برخی از دامین ها و یا اثرات مشاهده شده، مختص یک خانواده پروتئینی کامل است در صورتیکه دیگر نتایج مختص یک دامین است. به هنگام دستیابی به تعداد زیادی دامین که همگی یک چیز را توصیف می کنند، بانک های اطلاعاتی Interpro آنها را بصورت گروههایی در یک بخش جمع آوری می نماید، دقیقاً مشابه آنچه که در مورد دامین IPR004827 مشاهده می کنید.
- ✓ لینک IPR#### نشاندهنده مستندات موجود در Interpro است. در این بخش Interpro تلاش می کند تا خلاصه ای از اطلاعات بدست آمده از مستندات بانک های اطلاعاتی گوناگون دامین را ارائه نماید. مطالعه ی مستندات بدست آمده از Interpro ایده ی خوبی است، حتما تمامی مستندات مربوط به هر یک از توالی ها را به خوبی مطالعه نموده اید.
- ✓ در مقابل هر خط یک لینک برای دستیابی به ورودی دامین در بانک اطلاعاتی وجود دارد. برای مثال در صورت کلیک بر روی PS00138 Link، پس از جستجو، ورودی 138 از بانک اطلاعاتی prosite در اختیارتان قرار می گیرد، در این بخش شما می توانید به مستندات prosite دست یابید.
- ✓ بر روی هر خط، جعبه های کوچک رنگی، نشان دهنده ی محل تطابق ایجاد شده بر روی توالی مورد نظرتان است. اندازههای آنها دقیقاً بر اساس اندازه دامین می باشد. اگر نشان گر موس را بر روی مربع قرار دهید مختصات هر یک از تطابقات در توالی مورد نظرقابل مشاهده است.



شکل ۱۴-۶: خروجی بدست آمده از سرور Interproscan

درک صحیح از ناسازگاری ها و ناهماهنگی های احتمالی در خروجی بدست آمده از InterProScan

همانطور که در شکل ۱۴-۶ مشاهده می نمائید، تمامی بانک های اطلاعاتی دامین با وجود یک سری Leucine zippers که بطور کلی در وسط توالی مربوطه قرار گرفته، موافقند. از طرف دیگر، بانک های اطلاعاتی دامین بطور مشخصی با مرزهای این تطابقات هم عقیده نمی باشد. تشخیص درستی یا نادرستی مطلب نیاز به کمی مهارت دارد، با این حال در این مورد رأی اکثریت در نظر گرفته می شود (اکنون دلیل استفاده از تمامی مجموعه های دامین را دانستید و اینکه همه ی آنها مهم تلقی می شوند).

دیگر دلیل اینکه ۷ بانک اطلاعاتی بهتر از یکی است این است که همیشه این شانس وجود دارد که یک بانک اطلاعاتی حاوی دامین ای باشد که در دیگر بانک اطلاعاتی وجود ندارد. به خصوص هنگام مواجهه با دامین های جدید که هنوز در تمامی مجموعه های دامین وارد نشده اند کاربرد دارد. البته این حقیقت که یک دامین گزارش شده یا خیر همواره ۱۰۰ درصد اثبات همه چیز نیست (به بخش اشتباهات معمول به هنگام اسکن دامین ها مراجعه نمائید).

برای اطمینان از اینکه برخورد گزارش شده حقیقی است، پیشنهاد می کنیم نگاهی دقیق به هم ترازی موجود میان پروفایل و توالی مورد نظرتان بیاندازید. البته گزارش و یا عدم گزارش یک دامین لزوماً ۱۰۰٪ صحیح نمی باشد. متأسفانه، برنامه ی Interpro قادر به انجام چنین کاری نیست، باید به سراغ دیگر سرورها بروید (مانند CD- search و Motif- scan)که ادامه این بخش به توضیح آنها خواهیم پرداخت.

یافتن دامین ها از طریق سرور CD

سرور CD (دامین فشرده شده[1]) از NCBI ، در حقیقت از همان اصولی تبعیت می کند که برنامه Interpro از آن پیروی می کند و بر اساس آن ساخته شده است. مزیت اصلی این برنامه گزارش برخوردهای صورت گرفته با استفاده از درجه بندی آنهاست که در تشخیص و جدا سازی تطابقات خالص و حقیقی از ساختگی کمک می کند. از معضلات موجود این است که سرور CD حاوی بانک های اطلاعاتی به گستردگی Interpro نمی باشد با این حال تعداد اندکی دامین که از طریق NCBI منتشر شده است را شامل می شود این دامین ها را در هیچ کجای دیگر نخواهید یافت. در صورت تمایل به کار با این برنامه مطابق دستورالعملی که در ادامه آورده شده است رفتار نمائید:

۱. به آدرس اینترنتی http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi بروید. همچنین می توانید از طریق صفحه اصلی NCBI blast نیز به این سرور دسترسی پیدا کنید. برای این منظور به آدرس اینترنتی www.ncbi.nlm.nih.gov/BLAST/ بروید و بر روی گزینه Search the Conserved Domain کلیک نمائید (این لینک سومین لینک در ستون سوم است).

اشتباهات متداول به هنگام اسکن دامین ها

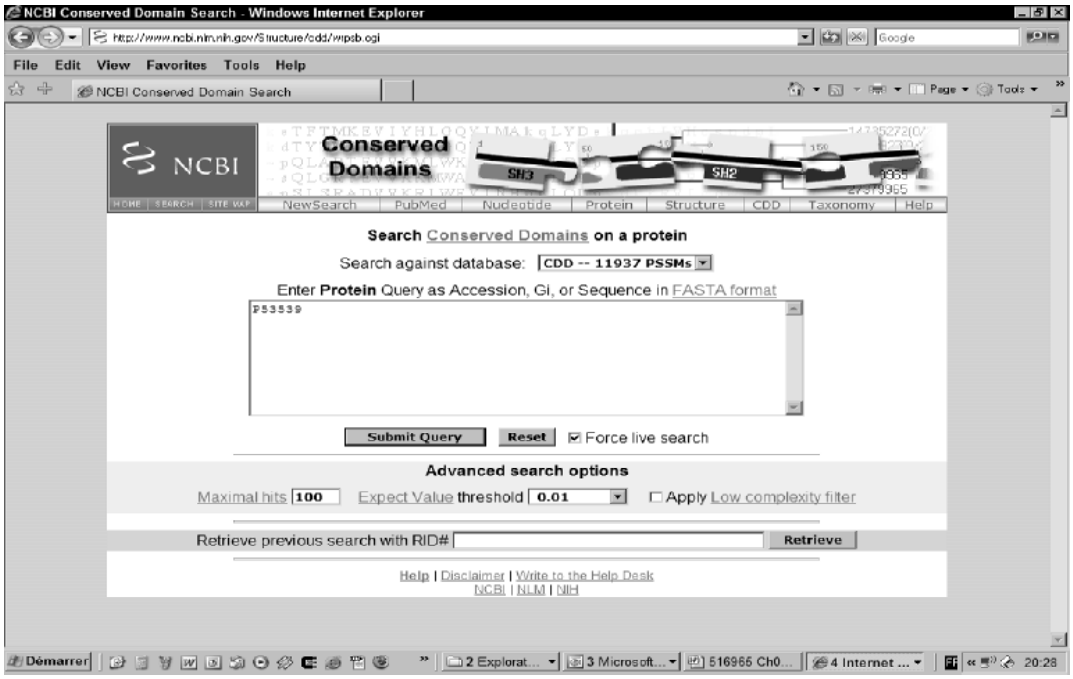
توجه نمائید که به هنگام گزارش انطباق مابین توالی مورد نظر و یک دامین از طریق یک سرور دامین، اطلاعات بدست آمده در حقیقت نتیجه ی هم ترازی های اعمال شده میان پروفایل دامین که شما یافته اید با توالی که از آن برای مقایسه استفاده می نمائید، می باشد. برنامه مربوطه در صورت تشخیص مناسب بودن هم ترازی اعمال شده آن را بصورت یک انطباق گزارش می نماید. در حالیکه ممکن است این طور نباشد بدبختانه نه پروفایل و نه برنامه خیلی مطلوب نیستند و ممکن است اشتباهاتی صورت گیرد. سرور مربوطه بیان گر وجود یک دامین خاص در پروتئین مورد نظرتان است در حالیکه این گونه نیست و یا بالعکس ممکن است دامین موجود در پروتئین شما را نادیده بگیرد. در صورتی که به این گونه نتایج بصورت کورکورانه اعتماد کنید ممکن است دست به اشتباهات بزرگی بزنید. سرور interpro scan خیلی هم برای جلوگیری از بروز این گونه مشکلات و اشتباهات مفید نمی باشد؛ دلیل آن عدم گزارش درجه بندی انطباق های احتمالی و نیز عدم نمایش هم ترازی های موجود است، دیگر سرورهای موجود که در این بخش به آن می پردازیم (motif search , CD-search) قادر به درجه بندی انطباقات ایجاد شده می باشند. این درجه بندی بر اساس مفاهیم آماری بوده و لذا بیانگر میزان ساختگی بودن و دخالت شانس در یک انطباق پیش آمده است. تفاسیر محافظه کارانه که بر اساس بالاترین درجه بندی موجود پایه ریزی می شوند در اکثر مواقع صحیح هستند. از طرف دیگر، در صورت نداشتن یک درجه خوب در لیست موجود یا در صورتیکه تنها بخشی از دامین مربوطه با توالی شما انطباق داشته باشد، به ناچار باید در پی یافتن شواهد بیشتر برای اطمینان از صحت مطلب ارائه شده در این زمینه باشید.

۲. توالی مورد نظر و یا شناساگر آن را در بخش جستجو در بالای صفحه وارد نمائید برای مثال می توان از پروتئین FOSB-HUMAN استفاده نمود. در صورت تمایل برای استفاده از این توالی، کد P53539 را در بخش جستجو مطابق آنچه در شکل ۱۵-۶ نشان داده شده است وارد نمائید.

[1] Conserved Domain

۳. گزینه Apply low Complexity filter را (که در زیر گزینه جستجوی پیشرفته قراردارد) از حالت انتخاب خارج کنید. در بسیاری از دامین ها اسید آمینه ای خاص بیش از حد انتظار وجود دارد. برای مثال، تعداد زیادی لوسین در مناطق Lucine Zipper وجود دارد و یا تعداد زیادتر از حد انتظار گلیسن در دامین های غنی از گلسین وجود دارد و نظایرهمین اتفاق برای دیگر دامین ها نیز دیده می شود. وجود اسیدهای آمینه تکراری باعث سادگی توالی ها شده و لذا به این توالی ها به اصطلاح توالی هایی با پیچیدگی کم (Low complexity) گفته می شود. در صورت انتخاب گزینه Apply low Complexity Filter ممکن است این دامین ها را از دست بدهید.

۴. گزینه Expect Value Threshold را ۱ انتخاب کنید. عدد پیش نویس (۰/۰۱) می تواند بیش از حد دقیق باشد بالاخص به هنگام در نظر گرفتن دامین های با پیچیدگی کم.



شکل ۱۵-۶: سرور جستجوی CD در NCBI

۵. بر روی گزینه Submit کلیک نمائید.

۶. نتایج خود را ذخیره نمائید. راحت ترین راه برای ذخیره سازی نتایج گرافیکی مورد نظر فشار دکمه Prt Sc و سپس Cut و Paste نمودن اطلاعات خود به درون PowerPoint می باشد.

تفسیر و درک صحیح نتایج حاصل از سرور CD

اصول حاکم بر خروجی های سرور NCBI CD بسیار شبیه به آنچیزی است که در Interpro با آن برخورد داشته اید. این اصول به دو بخش اصلی تقسیم می گردد:

- ✓ **تصاویر گرافیکی:** یک نمایش گرافیکی مطابق آنچه در شکل ۱۶-۶ مشاهده می نمائید در حقیقت نشان دهنده ی مناطقی از پروتئین مورد نظرتان است که با یک دامین انطباق دارد.
  - دامین های قرمز از SMART منشأ گرفته و

- انتهاهای برجسته (کنگره دار) نشان دهنده ی انطباق های نسبی است.

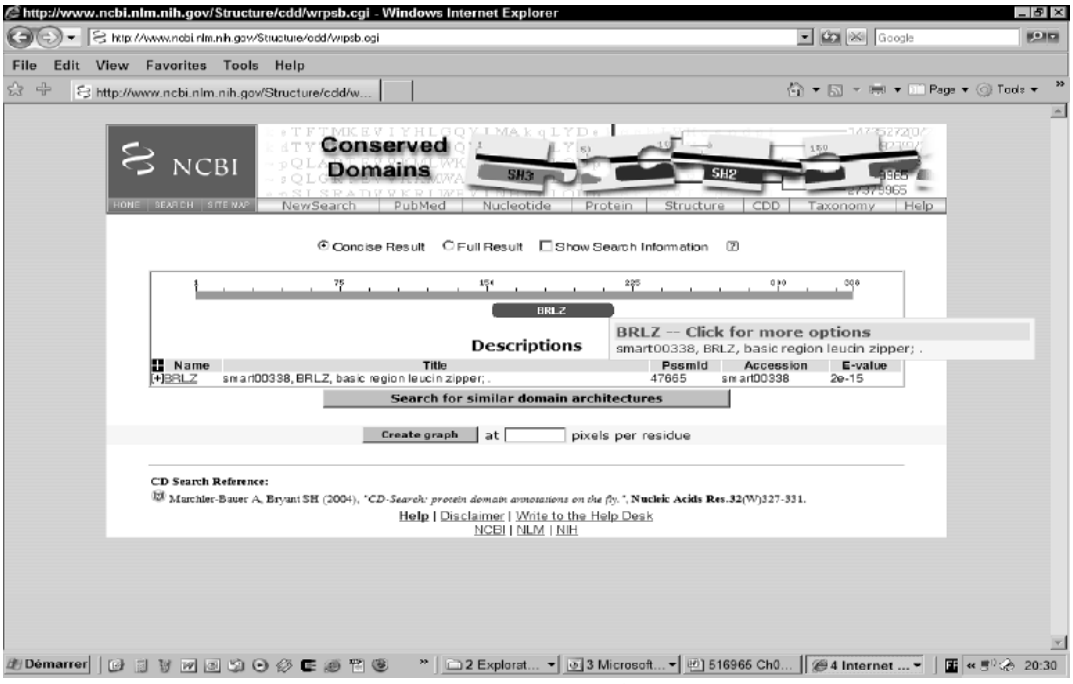
در صورت قرار دادن نشانگر موس بر روی یکی از دامین ها، نام کامل آن در پنجره ای کوچک نمایان می گردد. درجه بندی ارائه شده در حقیقت همان عدد E است که نشانگر تعداد دفعاتی است که انتظار می رودیک انطباق بر اساس شانس در توالی و یا بانک اطلاعاتی مورد جستجوی شما به وقوع بپیوندد.

برای تفسیر درجه بندی ارائه شده از قوانین زیر پیروی کنید:

- کمترین عدد E ارائه شده در حقیقت بهترین درجه بندی را داراست و نشان دهنده خلوص انطباق صورت گرفته است.
- تنها عدد E های زیر ۰/۰۱ معنادار می باشد.
- در صورت فیلتر نکردن دامین های با پیچیدگی کم میزان عدد E باید از ۰/۰۱ خیلی پائین تر باشد.
- انطباقات با انتهای برجسته را با احتیاط مورد بررسی قرار دهید غیر از مواردی که در مناطق با پیچیدگی کم اتفاق افتاده باشند، در مثال ارائه شده هیچکدام از انطباقات با انتهای برجسته بصورت بارز و آشکار وجود ندارد.

✓ **لیست انطباقات (The Hit List):** این لیست همانطور که در شکل ۱۷-۶ مشخص است، گزارشی از دامین هایی است با توالی مورد نظر انطباق داشته و بر اساس عدد E ذخیره سازی شده اند. بهترین انطباقات در ابتدای کار می آیند . در صورت کلیک بر روی هر یک از لینک های موجود صفحه ای شامل مستندات مرتبط با آن ورودی ظاهر می شود.

بر روی علامت (+) بعد از هر انطباق کلیک نمائید، با این کار هم ترازی های موجود مابین جستجوی شما و برخی از توای های دامین تخمین زده شده نشان داده می شود. این مختصات در حقیقت چیزی شبیه به میانگین توالی دامین می باشد.



شکل ۱۶-۶: خروجی گرافیکی بدست آمده از سرور NCBI CD

در صورت کلیک بر روی گزینه جستجو برای یافتن ساختارهای دامین مشابه (search for similar domain Architectures) در قسمت انتهایی لیست ارائه شده، سرور مربوطه تمامی توالی های پروتئین را که شامل دامین های مشابه با توالی مورد جستجو می باشد را در اختیارتان قرار می دهد. این توالی ها ممکن است بسیار مشابه باشند بطوری که از طریق راههای دیگر قابل شناسائی نمی باشند.

شکل ۱۷-۶: لیست انطباقات و همترازی های ارائه شده از طریق سرور NCBI CD

یافتن دامین ها با استفاده Motif Scan

افرادی که Motif scan را راه اندازی نموده اند در حقیقت همان کسانی هستند که در prosite به نوعی دست اندر کار بوده اند. این سرور اطلاعات prosite را با ظاهری قدرتمندتر ارائه می دهد. همچنین Motif scan شامل برخی از دامین هایی است که هر گز بطور رسمی از طریق Interpro ارائه نشده است.

در صورت نیافتن نتیجه ی مناسب در سرور Interpro scan و یا CD می توانید از این سرور استفاده نمائید. Motif Scan آخرین شانس شما برای یافتن اطلاعات مورد نظرتان است. مراحل زیر را دنبال کنید:

۱. به آدرس اینترنتی http://myhits.isb-sib.ch/cgi-bin/motif_scan بروید. صفحه ی اصلی سایت ظاهر می شود.
۲. توالی مورد نظر و یا کد شناسایی را در بخش ورودی توالی پروتئین وارد نمائید(مطابق شکل ۱۸-۶). شما می توانید از هر یک فرمت های زیر استفاده نمائید:
   الف) کد شناسایی توالی (Sequence ID)
   ب) فرمت خام به صورت اسیدهای آمینه که فواصل و اعداد نیز قابل بررسی است.
   ج) FASTA

برنامه Motif Scan بصورت اتوماتیک فرمت انتخابی را شناسائی می کند. برای مثال، می توانید از FOSB-HUMAN استفاده نمائید. برای این کار کافی است کد P53539 را در بخش ورودی توالی پروتئینی تایپ نمائید.

۳. در بخش بانک اطلاعاتی Motif (Database of Motifs)، مجموعه دامین مورد نظرتان را انتخاب نمائید.

PROSITE به نسبت Pfam کوچک تر بوده اما با انتخاب Prosite نتایج با سرعت بیشتری ارائه می شوند.

۴. بر روی گزینه جستجو کلیک نمائید. با انجام این کار صفحه ی نتایج بلافاصله ظاهر می شود.
۵. از روی نتایج خود پرینت بگیرید.

- ذخیره سازی نمایش گرافیکی: ابتدا بر روی Paste کلیک نمائید و سپس اطلاعات مورد نظر را به درون PowerPoint , Cut و Paste نمائید.
- ذخیره سازی نتایج: لیست انطباقات حاصل را Cut و Paste نمائید.

تفسیر و درک صحیح نتایج حاصل از برنامه Motif Scan

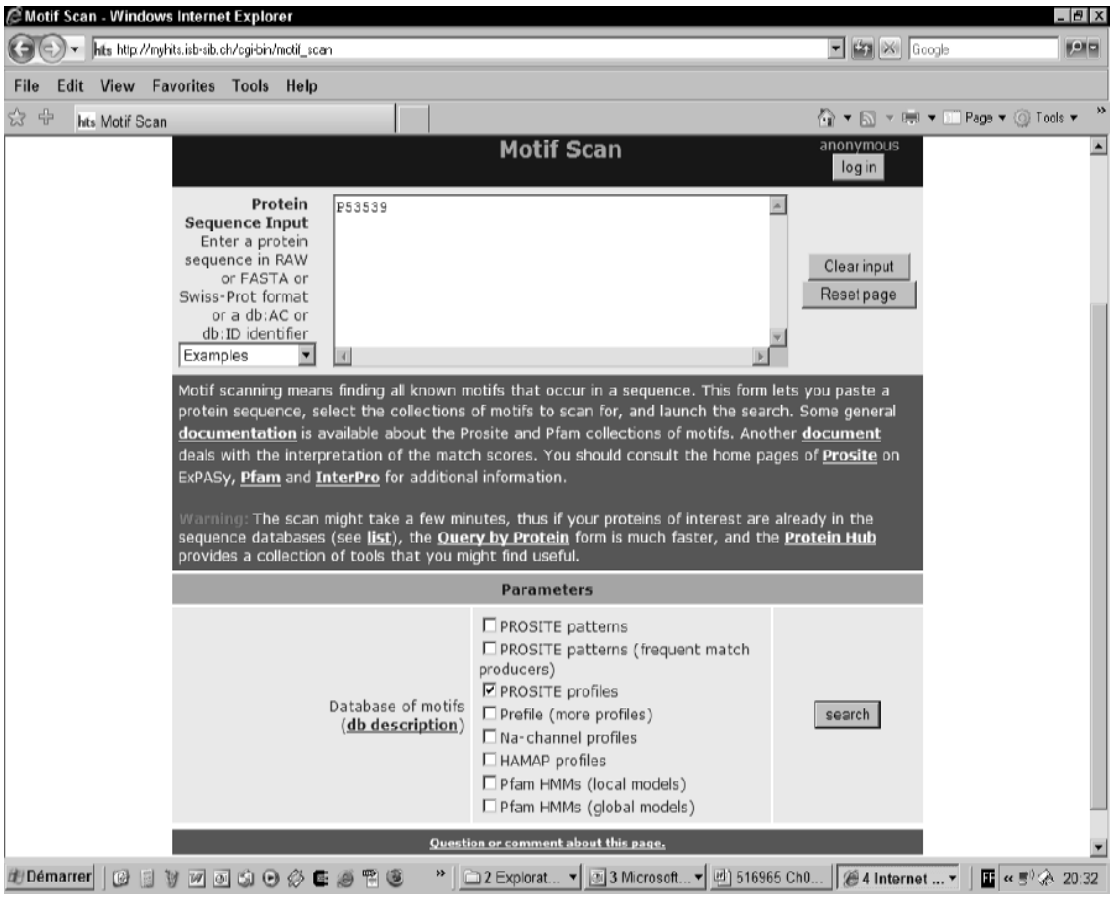
این برنامه یکی از غنی ترین خروجی های ممکن را از آنالیز دامین در اختیارتان می گذارد. این نتایج با ظاهری به مراتب دشوار تر از لحاظ تفسیر نسبت به CD و یا Interpro scan نمایش داده می شوند. برخی از نکات قابل ذکر برای تفسیر نتایج حاصل از motif scan به شرح زیر است:

- ✓ **Match map:** این نقشه نشان دهنده ی مناطقی است که هر دامین بر روی توالی مورد نظر منطبق شده است. هر یک از انطباقات صورت گرفته با یک عدد نشان داده شده و علامت اختصاری آن زیر هر انطباق قابل مشاهده است.
- ✓ **List of Matches:** این لیست نسخه متنی ارائه شده از نقشه ی انطباقات است. می توانید آن را کپی نمائید.
- ✓ **The Match Details Section:** این برنامه در حقیقت از یک درجه بندی معمول استفاده می نماید:
  - درجات بالا نشان دهنده یک انطباق خوب و مناسب است.
  - تنها درجات بالای ۷ نشان دهنده ی انطباق مناسب است.

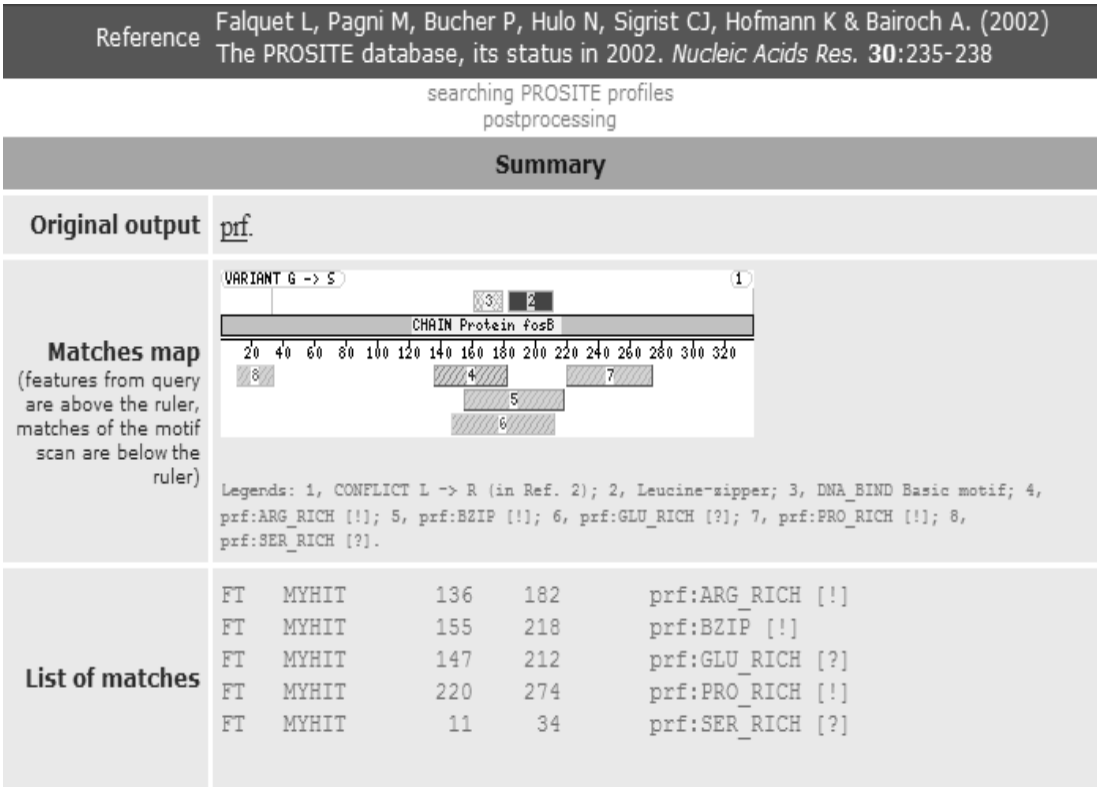
این برنامه نتایج بدست آمده را بر اساس درجات آنها ذخیره سازی نمی کند. بنابراین بهترین نتیجه لزوماً بالاترین مکان را در لیست ندارد، انطباقات نسبی با یک علامت تعجب (!) مشخص شده اند.

- ✓ **The Match Detail Representation:** برای هر انطباق گزارش شده Motif scan جزئیات انطباق (یکی از مشخصه های اختصاصی و منحصر به فرد این سرور، بخش *match detail*) را نشان می دهد. یک گراف (شکل گرافیکی) نشاندهنده این مطلب است که آیا اسیدهای آمینه ها در مکان های صحیح در توالی مورد نظر قرار دارند یا خیر. برای آگاهی از نحوه عملکرد آن نگاهی به ورودی مرتبط با B_ZIP Domain در شکل ۲۰-۶ بیاندازید.

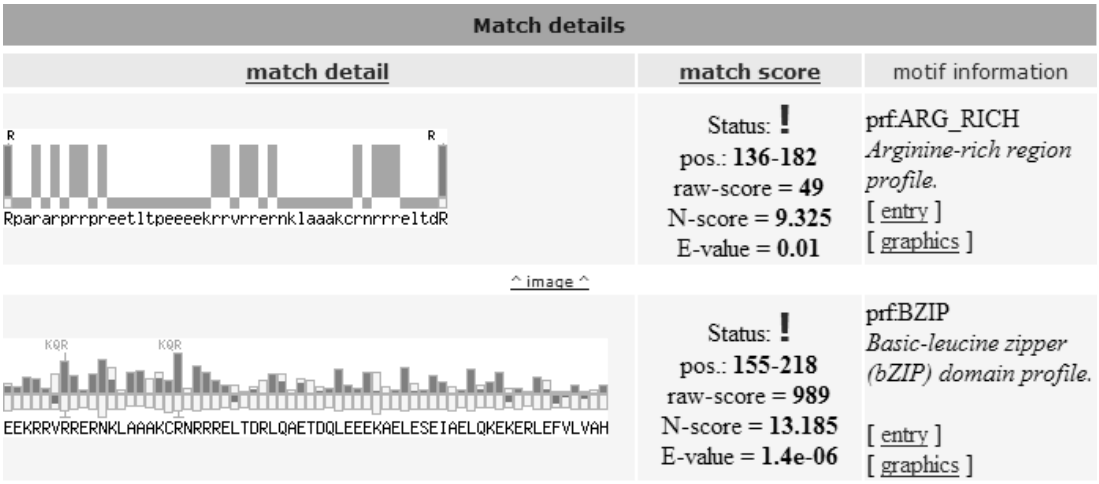
- ستونهای بالای هر توالی (رنگ سبز) نشان دهنده ی اسید آمینه هایی است که برای این دامین در این مکان محتمل تر می باشند. هر چه ستون مربوطه بلندتر باشد احتمال اختصاصی بودن آن برای اسید آمینه ی مربوطه بیشتراست و هر چه پهن تر باشد نشان دهنده این است که اسید آمینه مربوطه در مکان صحیحی بر روی توالی قرار گرفته است.
- ستونهای زیر هر توالی نشان دهنده ی مکانهایی است که توالی شما فاقد اسید آمینه صحیح است. طول ستون نشان دهنده احتمال عدم انطباق اسید آمینه مورد نظر در مکان صحیح است.



شکل ۱۸-۶: سرور Motif scan زیر نظر یک مؤسسه بیوانفورتیک سوئیس



شکل ۱۹-۶: خروجی حاصل از برنامه motif scan (قسمت فوقانی)



شکل ۲۰-۶: خروجی حاصل از Motif scan (قسمت انتهایی)

ستون بلند نشان دهنده آن است که اسید آمینه یافت شده نباید دراین مکان قرار گیرد (در جای صحیح قرار ندارد).

این نمایش در زمانیکه اسیدهای آمینه کلیدی (برای مثال جایگاههای فعال) در دامین مورد نظر قرار دارند، بسیار کارا است.

استنتاج صحیح از خروجی Motif Scan

در برنامه Motif scan، نمایش گرافیکی وسیله ای قدرتمند و کارا برای اطمینان از یک نتیجه گیری و برداشت صحیح می باشد. برای مثال، نمودار ارائه شده در شکل ۲۰-۶، می توان این گونه نتیجه گیری نمود که توالی مورد نظر شامل تمامی اسیدهای آمینه یک Basic-Leucine Zipper ،در مکان صحیح خود، می باشد.

در صورتی که یک اسید آمینه ی بسیار حفاظت شده (در بالاترین مکان ستون) با یک اسید آمینه غیر اختصاصی (در پائین ترین مکان ستون) جابجا شده باشد، در این صورت باید بسیار محتاط بود.

توجه نمائید، Motif scan همواره دو دامین را که از طریق سرور CD و یا Inter pro scan غیر قابل شناسائی هستند، در اختیارتان قرار دهند: دامین غنی از پرولین و غنی از آرژنین(شکل ۲۰-۶). با یک نگاه مختصر به توالی های مرتبط می توان به صحت دو دامین ارائه شده پی برد. این مطلب نشاندهنده یکی دیگر از مزایای استفاده از چندین سرور آنالیز دامین بجای یکی است. اکنون ما دو مطلب (مناطق غنی از آرژنین و پرولین) بیشتر در مورد توالی پروتئین خود می دانیم.

کشف دامین های جدید در پروتئین

یافتن دامین های جدید عملاً نیاز به کمی مهارت دارد که تنها توسط زیست شناسان آموزش دیده و با مهارت قابل انجام است. اما خوشبختانه تمام آنچه که شما می خواهید بدانید دردسترس است، البته در صورتیکه فقط می خواهید یک امتحان کنید. راحت ترین راه استفاده از BLAST و اختصاص دادن جستجوی بانک اطلاعاتی خود به آنچه BLAST به آن PSSM می گوید (و عموم به آن دامین)، می باشد. در صورتیکه با این ابزار آشنایی ندارید، بخش BLAST (فصل ۸) را حتماً مطالعه کنید. در این بخش تمام اطلاعات مورد نیاز شما برای ساخت و استفاده از PSSM ها در آدرس اینترنتی زیر در دسترس است:

www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm

در صورت استفاده از این روش، می توانید تمام اطلاعات را بصورت آنلاین در اختیار داشته باشید، اما انتظار معجزه را هم نداشته باشید. دامین ها به مانند یک الماس در دنیای پروتئین پراکنده شده اند. در صورتیکه گهگاه از روی شانس یکی از این گوهر های زیبا را بیابید بهتر از آن است که انبوهی از جواهرات را پیدا کنید. در صورت تمایل برای قدم گذاشتن در این راه نباید از کار با برنامه های Unix، جستجوی توالی های مورد نظر در بانک های اطلاعاتی عجیب و غریب توالی DNA (برای مثال بانک اطلاعاتی EST در بخش ۳ و ۴)، جمع آوری توالی های مورد نظر از طریق BLAST و نیز تنظیم و تطبیق آنها با برنامه Multiple Sequence Alignment (فصل ۹)، تغییر شکل هم ترازی بدست آمده بصورت مدل Hidden Markov Model (HMM) با استفاده از برنامه Hammer (hmmer.wustl.edu/) و نیز استفاده از برنامه Hammer برای جستجو در بانک های اطلاعاتی پروتئین نگران باشید، اینها کارهای معمولی است که در Pfam انجام می شود.

نقل است که اسکیموها برای کلمه ی برف ۴۰ واژه ی جایگزین دارند و تنها تفاوت این کلمات در کوچکترین تغییرات ایجاد شده است. مشابه این مثال برای زیست شناسان هم وجود دارد، اصلاً عجیب نیست که بشنوید بیش از یک واژه برای protein domains وجود دارد! در محتوای موجود در یک مقاله بیولوژیک، ممکن است از واژگانی چون *Profile, Pssm, HMM, MSA, Domain, Extended profile, Weight matrix* که همه تقریباً به یک معنا می باشند، استفاده نمائید.

روش های دیگر برای آنالیز پروتئین از طریق اینترنت

اینترنت شمار زیادی از منابع را برای انجام آنالیز آنلاین یک توالی ارائه می دهد و تمامی آنها رایگان هستند. تعدادی از آنها در جدول ۲-۶ بصورت خلاصه آورده شده است لینک سایت مربوطه در دسترس ترین راه برای آنالیز مورد نظر است.

بطور کلی در صورت کار با یکی از سایتهای مربوطه، پیشنهاد می کنیم از قابل اعتماد ترین منبع ارائه شده استفاده نمایند. به عنوان یک قانون هیچگاه از سایت های شخصی استفاده نکنید، زیرا بطور معمول مطالب آنها غیر قابل اعتماد می باشند.
www.something.somewhere/somebody

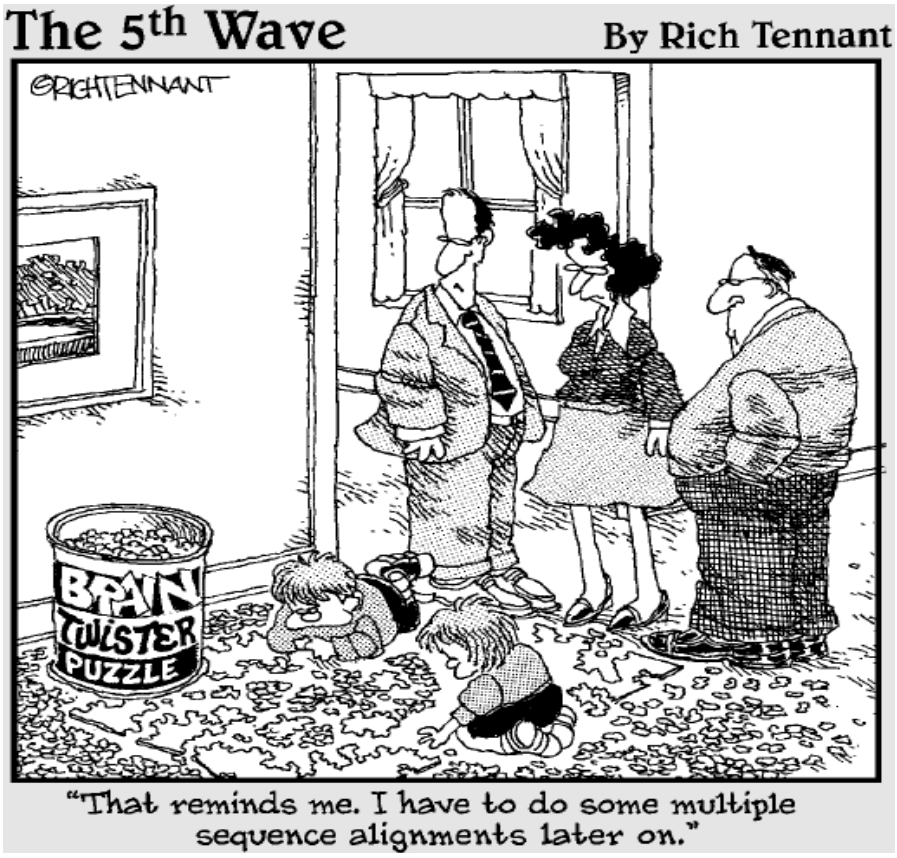
Table 6-2 Protein Sequence Analysis over the Internet

Name	*Site*	*Description*
ExPASy	www.expasy.org/tools	Proteins
Pbil	npsa-pbil.ibcp.fr	Proteins
PIR	pir.georgetown.edu	Proteins
CBS	www.cbs.dtu.dk/services	Proteins
Hits	hits.isb-sib.ch/	Proteins
InterPro	www.ebi.ac.uk/interpro/scan.html	Domains
CD search	http://www.ebi.ac.uk/InterProScan/	Domains

جدول ۲-۶ : آنالیز توالی پروتئین توسط اینترنت

بخش سوم

در آنالیز توالی ها مهارت کافی پیدا نمائید!



ترجمه: محمود حسینی

ویرایش: دکتر حسن کریکندی، سجاد عباسی

با تشکر از: فائزه زینلی، الهام دوست حسینی، ملک ناز باقیانی مقدم

در این بخش می خوانیم ...

اساس قدرتمند ترین ابزارهایی که در بیوانفورماتیک استفاده می شوند، روش های مقایسه توالی هاست. در این بخش به شما نشان می دهیم که چگونه یک بانک اطلاعاتی را با مقایسه توالی ها با استفاده از BLAST جستجو کنید، یا چگونه می توانید از تکنیک های مقایسه جفتی (مثل Dotlet) برای مقایسه دو توالی استفاده کنید.

همچنین در این بخش برنامه های multiple alignment را به شما معرفی می کنیم. این برنامه ها چندین توالی را همزمان با هم مقایسه می کنند. اگر با این موضوعات آشنا نیستید، بدانید که با دانستن آنها نحوه انجام تحقیقات بیولوژیک شما برای همیشه تغییر خواهد کرد!

فصل هفتم

جستجوی توالی های مشابه در بانک های اطلاعاتی

در این فصل می خوانیم:

- ✓ چرا تشابهات مهمند و چه ارزشی برای ما می توانند داشته باشند؟
- ✓ هومولوژی چیست؟
- ✓ چگونه از سیستم Replace & BLAST (Find) برای یافتن توالی های مشابه با یک توالی خاص استفاده کنیم؟
- ✓ چگونه بفهمیم که آیا نتایج به دست آمده از این جستجوها ارزشی دارند و یا نه؟
- ✓ بهترین نوع BLAST را نسبت به کاری که می خواهیم انجام دهیم انتخاب کنیم؟
- ✓ چگونه توالی های هومولوگ با شباهت کمتر را با BLAST–PSI جستجو کنیم؟
- ✓ چگونه مناطق جدید پروتیینی را با BLAST–PSI جستجو کنیم؟

.....

« یک آدم خوشبین، وقتی در انبار کاه به دنبال یک سوزن می گردد، دستکش هایش را به دست می کند»

اگر یک توالی پروتیین و یا DNA دارید و به دنبال یافتن توالی های مشابه با آن هستید، به بهترین منبع مراجعه کرده اید. ما در این فصل به شما یاد میدهیم که چگونه از سیستم BLAST برای یافتن توالی های مشابه و همچنین مقایسه ی آن ها با توالی مورد سوال خود استفاده کنید و تنها با دو کلیک موس به بهترین نتایج دست یابید. همچنین به شما نشان می دهیم که در چه موقعیت هایی بهتر است از BLAST–PSI (مدل جدید و قوی تر از BLAST) برای یافتن پاسخهای بیولوژیک خود استفاده کنید.

به عبارت دیگر، وقتی تنها داده ی شما نام یک توالی یا حتی کمی بیشتر از نام باشد(یعنی می توانید در حد چند کلمه آن راشرح دهید" برای مثال پروتئاز و سرین موش")، اولین کاری که باید انجام دهید این است که توالی اسیدآمینه ای ژنوم مورد نظر خود را از یک منبع اطلاعاتی به دست آورید. ما در بخش ۲ نحوه ی یافتن یک توالی با استفاده از نام آن را به شما یاد داده ایم.

چرا تشابهات مهمند؟!

معمولا توالی های مشابه از اجداد مشترکی مشتق شده اند. این بدان معنی است که توالی های مشابه احتمالا دارای اجداد مشترک، ساختمان مشترک و همچنین عملکرد بیولوژیکی مشترک هستند. این حقیقت حتی در مورد توالی هایی که از ارگانیسم های متفاوتی به دست آمده اند نیز می تواند صحت داشته باشد. به عبارتی دیگر اگر شما اطلاعاتی(اعم از ساختمان، عملکرد و ...) در مورد توالی خاصی داشته باشید می توانید همین اطلاعات را در مورد توالی های مشابه با آن بسط دهید.

برای مثال: تصور کنید دانشمندی مدتها وقت صرف پژوهش پیرامون توالی خاصی کرده و تمام داده های مربوط به آن را استخراج کرده است. اتفاقا این توالی به توالی مورد سوال شما شباهت زیادی دارد. بنابراین لازم نیست شما تمام مراحلی که او صرف کرده است را دوباره طی کنید بلکه تقریبا با اطمینان می توانید یافته های او را در مورد توالی خود صادق بدانید.

تصور کنید با دانستن این نکته چقدر در وقت خود صرفه جویی کرده اید. مطالعه روی یک ژن در آزمایشگاه سال ها زمان می برد در حالی که جستجوی یک داده در پایگاه های اطلاعاتی تنها به صرف چند ثانیه وقت نیاز دارد. با این کار بدون اینکه کسی شما را متقلب بداند روزها از وقت خود را صرفه جویی کرده اید.

بیولوژیست ها چنین توالی های مشابهی را « homologues» می نامند. به عبارتی وقتی واژه ی homologues در توصیف ۲ توالی بیان می شود، می توان ۲ توالی را فرض کرد که دارای اجداد مشترک و عملکرد و ساختمان مشابهی هستند. حالا چگونه می توانیم به دو توالی نسبت homologues را بدهیم؟ فرض کنید شما توالی با ۱۰۰ آمینواسید دارید. اگر توالی شما مربوط به یک پروتیین است طبق قوانین موجود، ۲۵٪ شباهت به بالا به معنای همولوگ می باشد. ولی اگر توالی شما مربوط به یک رشته DNA می باشد ۷۰٪ تشابه به بالا به معنای همولوگ است. بنابراین اگر عددی که به دست آورده اید کمتر از این مقادیر بود یعنی حدس شما در مورد همولوگ بودن دو توالی اشتباه بوده است." Rod Serling "چنین توالی هایی را مناطق « تیره و روشن (TWILIGHT ZONE) » نامید. بنابراین:

- ✓ حتی در صورتی که شباهتهایی(ساختاری، عملکردی و ...) میان دو توالی دیده شود، همولوگ بودن اثبات نمی شود. برای مثال پروتیین هایی وجود دارند که تشابه آمینواسیدی آن ها کمتر از ۱۵٪ می باشد ولی ساختار سه بعدی یکسانی دارند. در عین حال پروتیین هایی هم وجود دارند که تشابه آمینواسیدی آن ها ۲۰٪ بوده ولی هیچ شباهت ساختاری ندارند. طبق قوانین هیچ کدام از دو مورد ذکر شده همولوگ محسوب نمی شوند.
- ✓ چسباندن صفت همولوگ و یا غیر همولوگ طبق قوانین و مقرراتی است.

این ۲۵٪ ی که به عنوان یک شاخص برای تعیین مناطق تیره – روشن معرفی شد، تنها یک مقیاس معمولی و ساده محسوب می شود. در واقع مسایل، کمی پیچیده تر است. در بیشتر موارد برای اینکه از همولوگ بودن دو توالی مطمئن شویم باید از اطلاعات دیگری هم استفاده کنیم.اطلاعاتی نظیر:

- ✓ عدد E (Expected value) که به مانشان می دهد شباهت میان توالی یافته شده و توالی مورد نظر ما چه مقدار بر اساس تصادف بوده است.
- ✓ طول قطعات متشابه بین دو توالی.
- ✓ الگوی حفظ آمینو اسید ها در بین نسل ها.
- ✓ تعداد جهش ها(حذف و اضافه شدن آمینو اسیدها).

"شباهت (Similarity)" و "هولوگوس" را با هم اشتباه نکنید. همولوگ یک ارتباط دو طرفه و یک صفت مطلق میباشد، ولی شباهت یک صفت قابل اندازه گیری است. به عبارت دیگر دو توالی یا همولوگ هستند و یا نه. ولی دو توالی چه همولوگ باشند و چه نه، می توانند درجات مختلفی از شباهت را داشته باشند. برای مثال هیچ گاه نمی توانیم بگوییم دو توالی ۴۰٪ همولوگ هستند(مثل اینکه هیچ گاه نمیتوانیم بگوییم یک زن ۴۰٪ حامله است) ولی می توانیم بگوییم دو توالی ۴۰٪ با هم مشابهند.

در ادامه ی فصل چگونگی تفسیر چنین داده هایی را می آموزید.

محبوب ترین منبع به دست آوردن داده: BLAST

۳۰ سال پیش یک بیولوژیست برای اینکه بتواند از داده های موجود نتیجه ی مورد نظر را به دست آورد، باید تمام توالی های موجود در یک بانک اطلاعاتی را بر روی یک کاغذ نوشته و بر دیوار اتاقش نصب میکرد، سپس ساعت ها برای مقایسه ی توالی مورد نظر خود با تمام توالی های نصب شده بر دیوار زمان صرف می کرد، تا شاید به نتیجه ی مطلوب دست پیدا کند. ولی امروزه با وجود میلیون ها توالی شناسایی شده چنین کاری بعید به نظر می رسد. خوشبختانه امروزه کامپیوتر امکان استفاده از BLAST(The Basic Local Alignment and Search Tool) که در جهت حذف مرحله ی زمان بر پیدا کردن توالی مشابه از روی دیوار طراحی شده است- را برای ما فراهم کرده است.

در این بخش ۲ راه آسان برای جستجوی BLAST را به شما نشان خواهیم داد و مهم تر از آن چگونگی تفسیر نتایج به دست آمده از این جستجو را به شما یاد می دهیم. در آخر بخش نیز تعدادی ایده برای چگونگی استفاده از BLAST در حل مسایل دیگری که ممکن است برایتان پیش آید ارائه می دهیم(البته غیر از درست کردن قهوه!).

BLAST توالی های پروتیینی

BLAST توالی های پروتیینی یعنی کارهایی که برای به دست آوردن اطلاعاتی در مورد توالی های مشابه با توالی مورد نظر خود از بانک اطلاعاتی انجام می دهید. دو نوع از انواع BLAST که در رابطه با پروتیین وجود دارند :

- ✓ Blastp: یک توالی یک پروتیینی را با توالی های پروتیینی بانک اطلاعاتی مقایسه می کند.
- ✓ TBLASTn: یک توالی یک پروتیینی را با توالی های نوکلئوتیدی بانک اطلاعاتی مقایسه می کند.

اگر نمیدانید که از کدام نوع BLAST در مورد توالی پروتیینی خود باید استفاده کنید از Blastp استفاده کنید. در واقع هرگاه در انتخاب نوع BLAST مردد بودید، Blastp انتخاب برتر است. ولی باز هم اگر در برگزیدن Blastp به عنوان انتخاب بهتر مردد بودید، بهتر است برای خودتان مشخص کنید که هدف نهایی شما از این جستجو یافتن چه اطلاعاتیست، تا انتخاب راه برایتان آسان تر باشد. برای این کار می توانید از جدول ۱-۷ کمک بگیرید.

مسئله ی اصلی		انتخاب برتر
من اطلاعاتی پیرامون عملکرد توالی پروتیینی خود می خواهم.		از Blastp استفاده کنید تا پروتیین شما را با پروتیین های موجود در بانک اطلاعاتی خود مقایسه کند.
من می خواهم ژن های جدیدی که پروتیین های ساده ای را کد می کنند کشف کنم.		از tBLASTn استفاده کنید تا پروتیین شما را با هر ۶ چارچوب قابل خواندن توالی های DNA مقایسه کند.(۳ تا در هر رشته)

جدول ۱-۷

در این بخش ۲ مورد ازمعروفترین Blastp های آنلاین و نحوه ی استفاده از آن ها به شما معرفی می کنیم:

- ✓ NCBI(PubMed).
- ✓ Swiss EMBnet.

این دو سرور (Replace & Find) کمی با هم متفاوتند و اگر شما هر دوی آن ها را به خوبی یاد گرفتید می توانید مطمئن باشید در استفاده از سایر سرورهای معرفی شده در پایان فصل نیز توانایی کافی را دارید.

۲ دلیل خوب برای لزوم یادگیری استفاده از سرورهای متنوع BLAST وجود دارد:

- ✓ منابع اطلاعاتی متنوع: هر server به یک منبع اطلاعاتی خاص متصل می باشد. یعنی شما اگر از جستجو در یک server نتیجه ای به دست نیاوردید ممکن است در یک server دیگر اطلاعات مورد نیاز خود را بیابید.
- ✓ سرعت: ممکن است یک server به علت هجوم همزمان کاربران، سرعت پایینی داشته باشد. در این مواقع استفاده از یک server دیگر میتواند انتخاب مناسبی باشد.

به علت میزانی از تفاوت در«نسخه منبع اطلاعاتی» server های گوناگون، بسیار به ندرت اتفاق می افتد که با دادن یک داده ی مشترک به ۲ server متفاوت نتایجی کاملا مشابه به دست آید. البته تفاوتها خیلی چشمگیر نخواهند بود ولی بنابر دلیلی که گفته شد اندکی تفاوت طبیعی است.

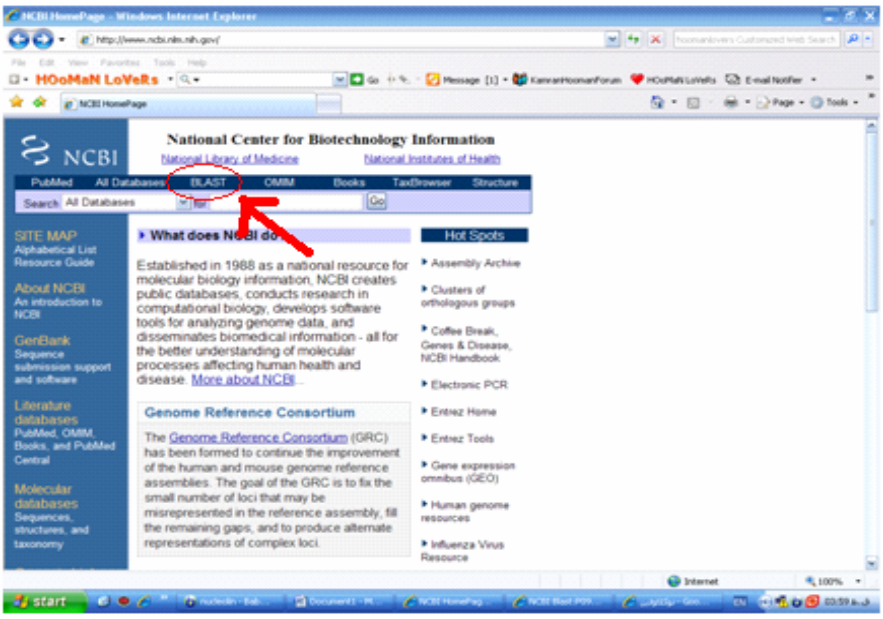
قبل از اینکه شروع به توضیح نحوه ی استفاده از سرورهای نام برده کنیم لازم است سوالی به عنوان پیش فرض مطرح کنیم و سپس با استفاده از مطالبی که یاد میگیریم جوابی مناسب برای آن بیابیم.

کار کردن با NCBI Blastp

سوال مورد نظر ما این است: « آیا پروتیینی مشابه با نوکلئولین (نوعی pro موجود در ماده ی سازنده ی هستک) هامسفر در پایگاه داده ی Swiss-Prot وجود دارد یا نه؟»

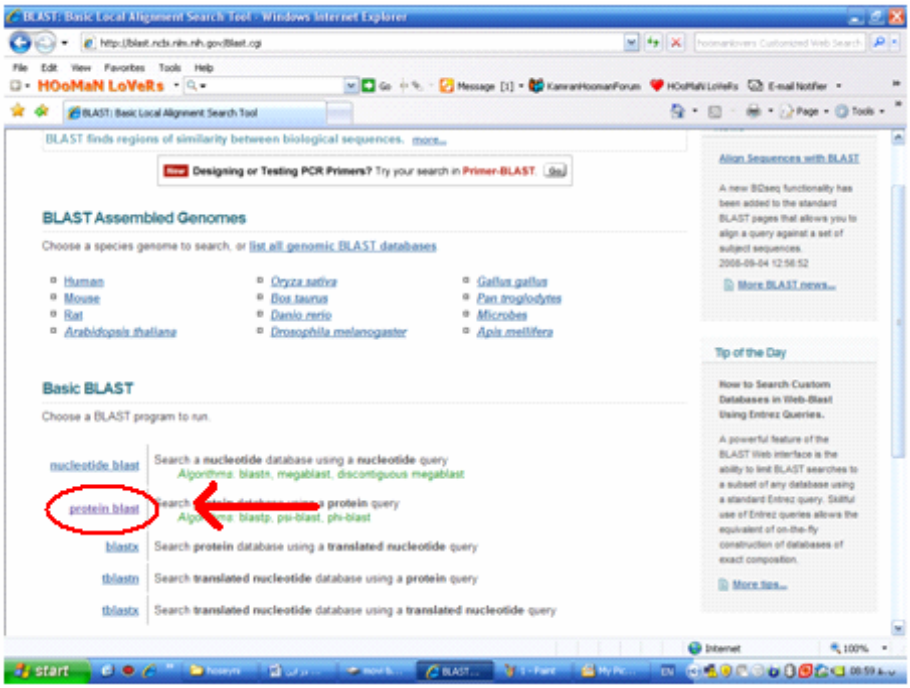
حال به دنبال جواب می رویم:

۱. وارد سایت www.ncbi.nlm.nih.gov شوید. صفحه ای مشابه صفحه ی زیر باز می شود.



شکل 7-1

همان طور که در شکل 7-1 میبینید، در کادر آبی بالای صفحه قسمتی به نام BLAST وجود دارد. بر روی BLAST کلیک کنید. صفحه ای زیر مطابق شکل ۷-۲ باز می شود.



شکل ۷-۲

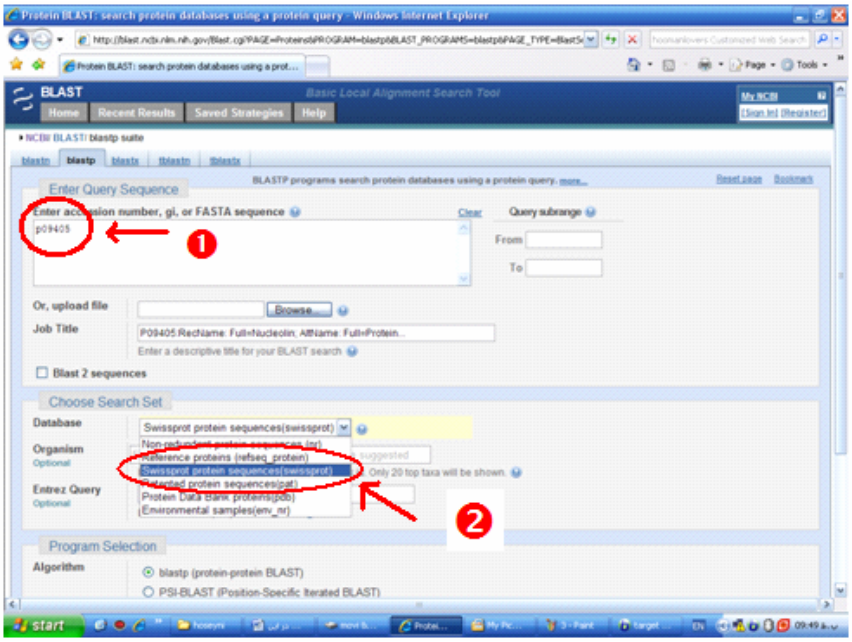
۲. طبق شکل ۷-۲ بر روی BLAST– protein کلیک کنید. صفحه ای مشابه با شکل ۷-۳ باز می شود. در این صفحه شما باید اطلاعات خواسته شده را وارد کنید.

۳. نوبت به پر کردن جعبه جستجو رسید. بسته به میزان دانسته های شما پیرامون سوال مورد نظر روش جستجوی شما نیز تغییر میکند. بدین معنی که ممکن است:

- ✓ توالی مورد سوال شما یک توالی قدیمی باشد که قبلا برای بانک های اطلاعاتی تعریف شده و به تبع آن یک کد شناسایی به آن اختصاص داده شده است و شما این شماره را میدانید و یا
- ✓ ممکن است شما خود، توالی خاصی را کشف کرده اید که قبلا وجود نداشته و یا اینکه توالی برای بانک اطلاعاتی تعریف شده است ولی شما کد شناسایی آن را نمیدانید و تنها داده ای که در دست دارید فرمت FASTA ی آن می باشد.

به هر صورت اگر کد شناسایی را میدانید در قسمت مشخص شده وارد کنید و در غیر آن صورت در همین قسمت فرمت FASTA آن را وارد کنید.

« p09405 کد شناسایی پروتیین مورد نظر ما است، بنابراین این عدد را در جعبه جستجو وارد می کنیم.(شکل ۱-۳)



شکل ۳-۷

توالی که در جعبه جستجوی Blastp وارد کردید همان توالی مورد سوال بود.

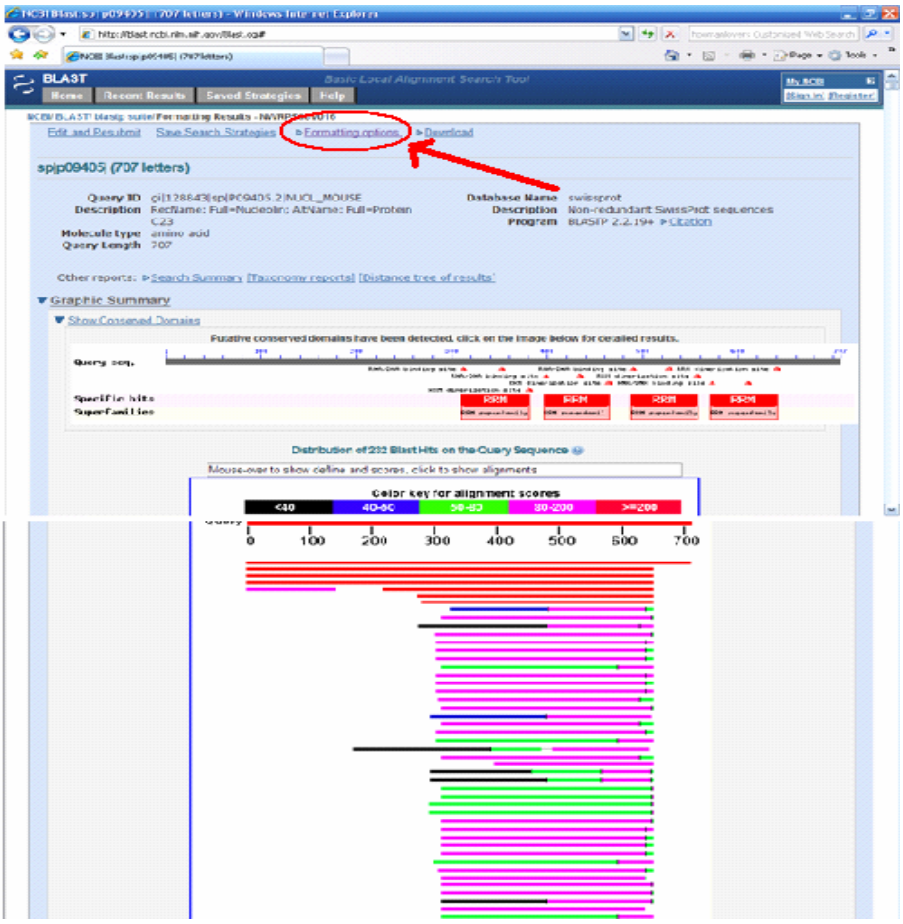
توالی هایی که Blastp به شما خواهد داد همان توالی های مشابهی است که در پی آن بودید.

بانک اطلاعاتی که در آن جستجو می کنید همان بانک اطلاعاتی هدف می باشد.

۴. طبق سوال مطرح شده، ما می خواهیم پاسخ سوال ما از پایگاه داده ای Swiss-Prot باشد. بنابراین از قسمت Choose Search Set و در قسمت Database گزینه ی Swiss-Prot protein sequence را انتخاب می کنیم.(شکل ۳-۲)

۵. مربع کنار گزینه Do CD Search را تیک بزنید.

۶. بر روی گزینه ی BLAST که در پایین صفحه وجود دارد کلیک کنید و منتظر بمانید. صفحه ای باز می شود که شامل تمام اطلاعات مورد نیاز شما می باشد.



شکل ۴-۷

Query subrange: شما به کمک این قسمت می توانید یک قسمت کوچک از یک توالی بزرگ را مورد بررسی قرار دهید. به عنوان مثال می خواهیم توالی های مشابه با مناطق ۲۴ تا ۲۰۰ یک توالی خاص را بیابیم. برای این کار، در کادر from عدد ۲۴ و در کادر to عدد ۲۰۰ را وارد می کنیم.

BLAST ۲ sequence : به شما این امکان را میدهد که ۲ توالی خاص را با هم مقایسه کنید.

به طور معمول باز شدن صفحه ی حاوی نتایج جستجو در Swiss-Prot یا NR تنها چند دقیقه زمان می برد ولی گاهی پیش می آید موتور جستجوگر این اخطار را به شما می دهد که دسترسی به نتیجه ممکن است بیش از ۲۰ دقیقه طول بکشد. در این زمان

شاید بهتر باشد که شما یک بار دیگر روند ذکر شده را از ابتدا اجرا کنید و یا اینکه در یک سرویس BLAST دیگر شانس خود را امتحان کنید.

در مدتی که منتظر باز شدن صفحه ی حاوی نتایج هستید، بر روی هیچ کلید و یا گزینه ای کلیک نکنید!

اگر نتیجه ای باب میل شما از یک موتور جستجو به دست نیامد، تکرار مجدد جستجو در جهت جوابی برای سوال قبلی، نتیجه ی جدیدی در بر نخواهد داشت. پس مجددا توالی خود را در جعبه جستجو وارد نکنید!

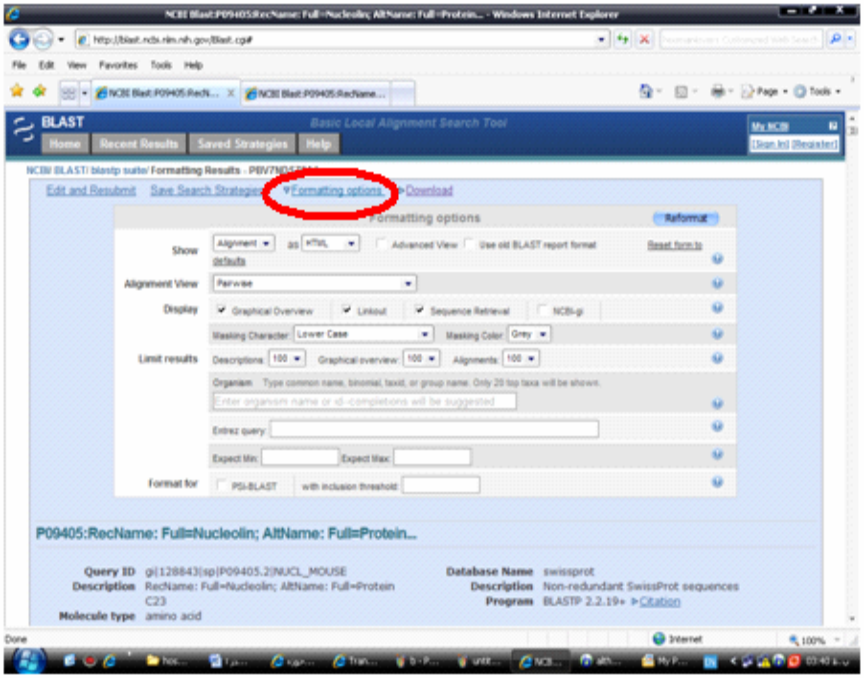
۷. مرحله ی آخر ذخیره ی اطلاعات به دست امده است.

هیچ گاه برای صفحه ی نتایج دستور پرینت ندهید، چون قسمت Alignment صفحه خیلی طولانی است. برای ذخیره ی این صفحه تنها کافی است از منوی File گزینه ی As Save را انتخاب کنید.

متأسفانه با روش قبل تصاویر گرافیکی ذخیره نخواهند شد. برای ذخیره ی آن ها ۲ راهکار وجود دارد:

- ✓ ذخیره ی تصاویر گرافیکی: بر روی تصاویر مورد نظر راست کلیک کرده و گزینه ی Save picture as را انتخاب کنید. تصویر با فرمت GIF ذخیره خواهد شد. گرچه اگر خواستید می توانید بعدا آن را به بقیه ی اطلاعات ذخیره شده اضافه کنید.(به یاد داشته باشید که تصویر دیگر فعال نخواهد بود.)
- ✓ نتایج را به فرمت PDF ذخیره کنید: بسته به نوع برنامه ی PDF ی که استفاده می کنید ممکن است تصاویر فعال باقی بمانند و یا نمانند.

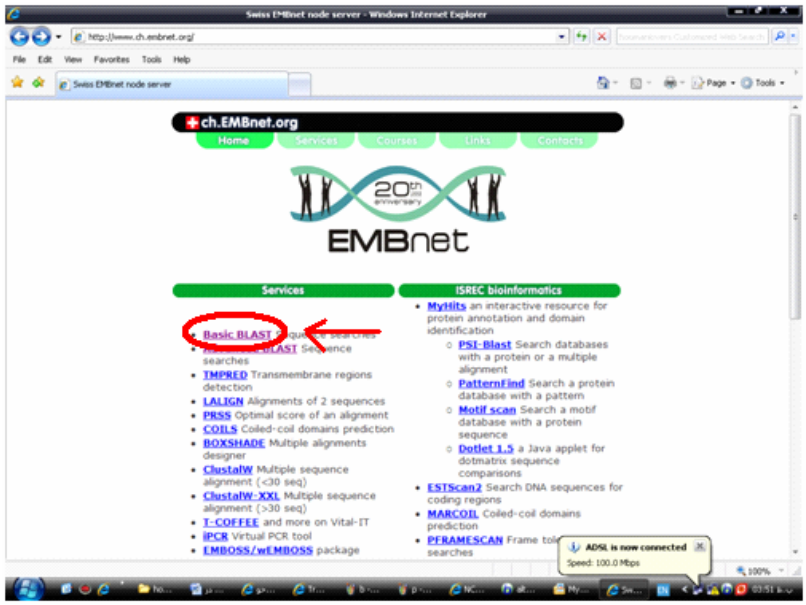
در بالای صفحه ی نتایج گزینه ای به نام Formatting option وجود دارد(شکل ۴-۷). با کلیک بر روی آن صفحه ی کوچکی باز می شود که می توانید با تغییر در پارامترهای آن نحوه ی نمایش نتایج را به دلخواه تنظیم کنید.(برای مثال تعداد توالی های نمایش داده شده، نشان دادن و یا ندادن Alignment ها و .....)(شکل ۵-۷)



شکل ۵-۷

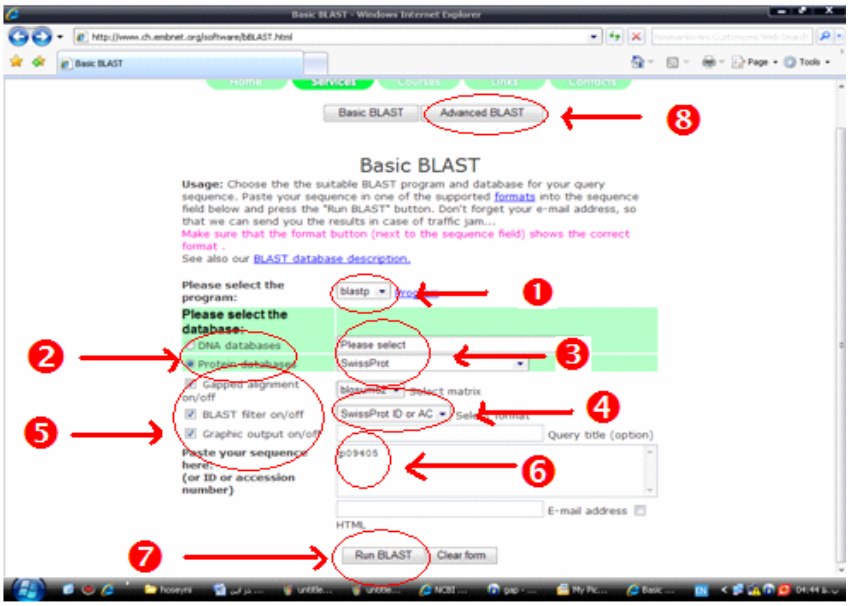
کار با موتور جستجوگر EMBnet blastp

۱. آدرس اینترنتی www.ch.embnet.org در نوار جستجو وارد کنید تا صفحه ی اصلی سایت باز شود(شکل 7-6) .



شکل ۶-۷

۲. در صفحه ی باز شده بر روی گزینه ی BLAST Basic کلیک کنید.(شکل ۶-۷) صفحه ی مربوطه باز می شود.(شکل ۷-۷).



شکل 7-7

۳. از منوی program گزینه ی blastp را انتخاب کنید (شکل ۷ – ۷).

این منو در واقع به شما این امکان را میدهد که سرویس BLAST مورد نظر خود را به راحتی انتخاب کنید.

۴. در مرحله ی بعد باید از میان دو عنوان پیشنهادی DNA databases و Protein databases گزینه ی متناسب با جستجوی خود را انتخاب کنید.

انتخاب بانک اطلاعاتی (databases) باید متناسب با نوع BLAST انتخابی در مرحله ی ۲ باشد. متأسفانه موتور جستجوگر EMBnet انتخاب های غیرممکن را حذف نمیکند و خود شا باید از میان گزینه های موجود، گزینه ی مناسب را انتخاب کنید. برای این کار شاید جدول زیر بتواند به شما در انتخاب درست کمک کند:

BLAST انتخابی	پایگاه داده ای متناسب
Blastp, BLASTx	Protein
BLASTn, tBLASTn, tBLASTx	DNA

بر اساس این جدول انتخاب متناسب با سوال مورد نظر ما Protein databases می باشد.(شکل ۲-۷)

قدم بعدی انتخاب یکی از گزینه های موجود در منوی مقابل Protein databases است.(البته روشن است که اگر انتخاب ما DNA databases بود، در این مرحله هم باید از منوی مقابل آن گزینه ای را انتخاب می کردیم!)(شکل ۳-۷)

تجزیه و تحلیل نتایج به دست آمده از یک « پایگاه داده ی ژنومی » بسیار راحت تر از بقیه می باشد.بنابراین در مواقعی که توالی مورد نظر شما مربوط به یک گونه ی بزرگیست، استفاده از این بانک های اطلاعاتی دست یابی به نتیجه را راحت تر خواهد کرد، البته به شرطی که شما نام گونه را بدانید. ولی اگر حتی نام گونه را نمیدانید، نگران نباشید. در بالای همین صفحه بر روی گزینه ی BLAST Advanced (شماره ۸-۷)کلیک کنید و گونه ی ژنومی مورد نظر خود را از صفحه ی جدید انتخاب کنید. برای شناسایی آن ها تنها کافی است به نامشان دقت کنید (نامشان متشکل از نام اختصاصی گونه + کلمه ی Genome می باشد). نگاهی به منوی مقابل DNA databases بیندازید، نمونه هایی از آن را خواهید دید.(مانند C.Elegans Genome).

اگر بانک اطلاعاتی مورد نظر شما در این جا وجود نداشت، بهتر است از سرورهای دیگر ارائه دهنده ی BLAST استفاده کنید. به عنوان مثال سایت مؤسسه ی بیوانفورماتیک اروپا می تواند کمک کننده باشد« (www.ebi.ac.uk/blast/).

بر اساس پیش فرض اولیه ما که استفاده از Swiss –Prot بود، انتخاب ما در اینجا همان Swiss –Prot است.

۵. به مرحله ی تنظیم « Select format» رسیدیم. فرمت «Plain text» فرمتی مشابه FASTA در اختیار شما می گذارد، البته با یک تفاوت جزیی و آن اینکه خط اولی که با > شروع می شود را خود به خود حذف می کند.

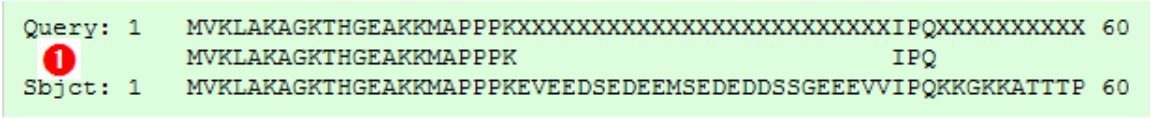
بقیه ی فرمت های موجود در این منو، داده ها را به صورت عدد در اختیار شما قرار می دهند.

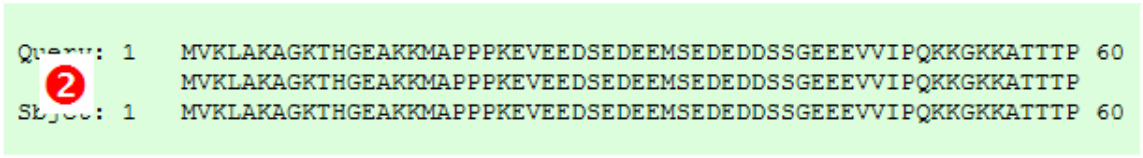
ما در اینجا Swiss-Prot or AC را انتخاب می کنیم (شکل شماره ۴-۷).

۳ مربع با سه عنوان بر روی صفحه قابل مشاهده است (شکل شماره ۵-۷):

- ✓ Gapped alignment on/off: در واقع با انتخاب این مورد شکل ارائه ی نتایج مشابه نتایج BLAST NCBI است. مربع کنار این گزینه را در حالت انتخاب قرار دهید.

✓ BLAST filter on/off: انتخاب این مورد باعث می شود مناطقی از توالی، که یک اسید آمینه بارها تکرار شده است حذف شود تا از میزان خطای ما کاسته شود. (شکل ۱-۸ On و شکل ۲-۸ off)





شکل ۸

✓ Graphic output on/off: نمودارهایی مشابه با نمودارهای BLAST NCBI را برای شما ترسیم می کند.

این نوار را انتخاب کنید.

۶. توالی مورد نظر و یا کد شناسایی توالی مورد نظر خود را در جعبه مشخص شده بنویسد.

ما در اینجا p09405 را وارد میکنیم. (شماره ۶-۷)

۷. اکنون به مرحله ای رسیده ایم که میتوانیم به نتایج مورد نظر دست یابیم. برای رسیدن به نتیجه بر روی BLAST Run که در پایین صفحه وجود دارد کلیک کنید (شماره ۷-۷).

انتخاب گزینه ی clear form تمام اطلاعات وارد شده را پاک میکند.

۸. برای ذخیره کردن نتایج به دست آمده میتوانید از منوی file گزینه ی save را انتخاب کنید. در نظر داشته باشید که با این فرایند، نمودارهای ترسیمی save نخواهد شد. اگر به ذخیره ی آن ها نیز نیاز دارید می توانید بر روی هر کدام، راست کلیک کرده و از منوی باز شده گزینه ی save picture as را انتخاب کنید.

میزان نتایج حاصل از BLAST خیلی زیاد می باشد. بنابراین به یاد داشته باشید غیر از مواردی که به همه ی آن نتایج نیاز دارید هرگز به کامپیوتر دستور پرینت ندهید.

اگر توانایی ذخیره نتایج را به صورت فرمت PDF دارید، حتما به این صورت ذخیره کنید. PDF بهترین راه برای ضبط و ثبت داده های الکترونیک می باشد.

چگونه نتایج حاصل از جستجو در یک موتور BLAST را بفهمیم؟

نتایج حاصل از تمامی انواع BLAST (BLASTn, tBLASTn, tBLASTx و ...) از نظر کلی با هم مشابهند و تفسیر آن حتی برای یک فرد باتجربه نیز پیچیده و گیج کننده خواهد بود. نتایج به دست آمده شامل ۴ بخش کلی می باشند. اگر شما توانایی تفسیر همه ی این ۴ بخش را دارید می توان شما را یک دانشمند بیوانفورماتیک نامید و یا حتی می توان نتیجه گرفت که شما دارای یک نیروی خارق العاده هستید!

نمایی کلی از خروجی BLAST

در داده های خروجی اغلب سرورهای BLAST ، ۴ قسمت اصلی مشاهده می شود که شامل:

۱. نمودارهای ترسیمی(Graphic display): از این قسمت می توانید مناطق متشابه بین توالی خود و سایر توالی ها را بیابید.بسته به نوع موتور جستجویی که استفاده کرده اید نتایج می تواند بسیار متغیر باشد. این قسمت در بعضی موتورها تعریف نشده است و بنابراین در نتایج جستجوی خود در آن موتورها، این قسمت را نخواهید دید.

۲. لیست هدف (Hit list): در این قسمت تمام توالی های مشابه با توالی مورد نظر شما با ذکر نام بیان شده و بر اساس میزان شباهت مرتب شده اند.

۳. هم ترازی ها(Alignment): در این بخش هم ترازی بین توالی شماو توالی های مشابه یافته شده نشان داده شده است.

۴. پارامترها(parameters): لیستی از تمام پارامترهایی که بر اساس آن ها جستجو انجام گرفته است.

هر کدام از قسمت هایی که نام برده شد، حاوی گنجینه ای از اطلاعات می باشد که دانستن آنها در یافتن پاسخی صحیح برای سوال مورد نظر ما می تواند مفید باشد.

نمودارهای ترسیمی

شکل ۹-۷ نمونه ای از نمودارهای ترسیمی در صفحه ی نتایج BLAST می باشد. میله ی نمایش دهنده توالی مورد سوال در بالای این قسمت قرار گرفته است(شکل ۹-۷ شماره ۱). هر کدام از میله های رنگی که در زیر میله ی اصلی دیده می شوند، نماینده ی توالی های دیگر متشابه با توالی شما هستند. به وسیله ی این میله ها می توان محدوده متشابه، بین توالی های مشابه با توالی مورد سوال را نیزبا یک نگاه مشاهده کرد.

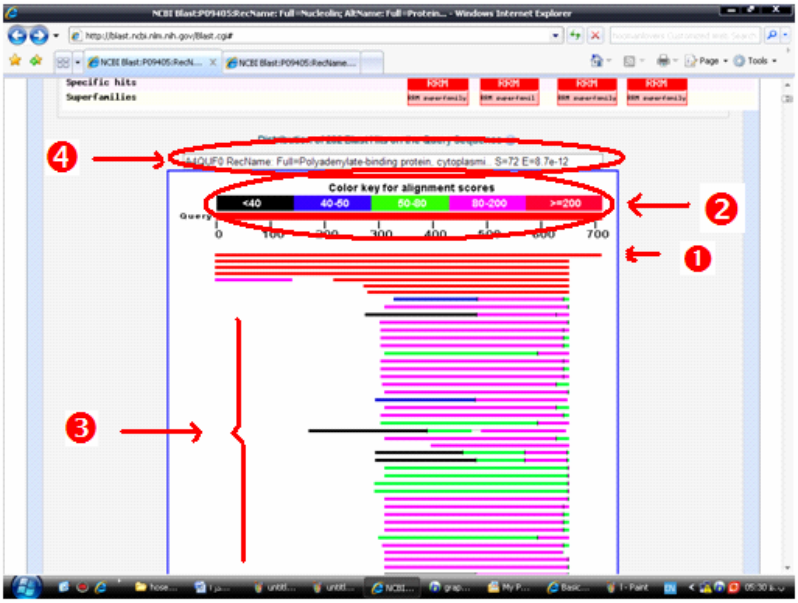
رنگ قرمز نشان دهنده ی توالی هایی است که بیشترین شباهت را با توالی شما دارند. رنگ صورتی، توالی هایی را نشان می دهد که میزان شباهت آن ها مقداری کمتر می باشد. رنگ سبز توالی هایی هستند که به اندازه دو رنگ قبلی مشابهت ندارند ولی میزان مشابهتشان خیلی پایین نیست و رنگ های آبی و سیاه هم توالی هایی را نشان می دهند که کمترین شباهت را به توالی مورد نظر دارند.(در بالای این نمودار قسمتی وجود دارد که به طور مختصر اطلاعاتی را در مورد چگونگی تفسیر رنگ ها در اختیارتان قرار می دهد (شکل ۷-۹ شماره ۲).

میله های قرمز، صورتی و سبز معمولا نشان دهنده ی توالی هایی با شباهت خوب هستند. توالی های نا مناسب پروتیین هایی هستند که کمترین شباهت را با توالی مورد نظر دارند و نمودار ترسیمی آن ها احتمالا هیچ شباهتی را از نقطه نظر بیولوژیکی نشان نخواهد داد.(در واقع این پروتیین ها همان مناطق تیره و روشن هستند)

نکته ی مثبت دیگری هم در طراحی این قسمت وجود دارد. به شکل ۷-۹ شماره ۳ نگاه کنید. اگر دقت کنید می بینید که بعضی از توالی های موجود، طول کمتری نسبت به توالی اصلی دارند. توالی هایی که در بالای صفحه قرار دارند و هم اندازه ی توالی اصلی هستند، در واقع پروتیین های همولوگ « نوکلئولین» هستند. ولی توالی های کوتاه تری که در پایین می بینید حاوی تنها یک منطقه آمینواسیدی مشابه با نوکلئولین هستند(یعنی یک منطقه از ژن نوکلئولین مشابه با منطقه ای از ژن این پروتیین ها می باشد). تنها شباهت میان نوکلئولین و این پروتیین ها،شباهت قسمتی از RNA این ژن ها با قسمتی از RNAنوکلئولین (به علت شباهت مقداری از ماده ی ژنتیکی)می باشد و هیچ شباهت دیگری بین آن ها نمی توان یافت.این خصیصه ما را در پیدا کردن دامین ژن ها (Domain)کمک می کند.

این نمودار های ترسیمی در NCBI فعال هستند. بدان معنی که اگر با موس روی میله های رنگی حرکت کنید، نام توالی مربوط به آن میله در مربع کوچکی که بالای همین تصویر هست نمایان می شود(شکل ۷-۹ شماره ۴) و حتی با کلیک بر روی هر کدام از میله های رنگی سرور شما را مستقیما به قسمت مربوطه در Alignment می برد.(البته دقت داشته باشید این نمودار در EMBnet چنین امتیازاتی را برای شما فراهم نمیکند).

در بعضی از قسمت ها ممکن است دو میله رنگی با یک خط باریک سیاه به یکدیگر متصل شده و یک میله ی رنگی را تشکیل داده باشند. چنین میله هایی نشان دهنده ی پروتیینی هستند که در ۲ قسمت متفاوت آن، توالی مشابه با توالی پروتیین شما دارد.(چنین پروتیینی در EMBnet به جای خط سیاه باریک با یک خط تیره کوتاه نشان داده می شود).



شکل ۷-۹

لیست هدف

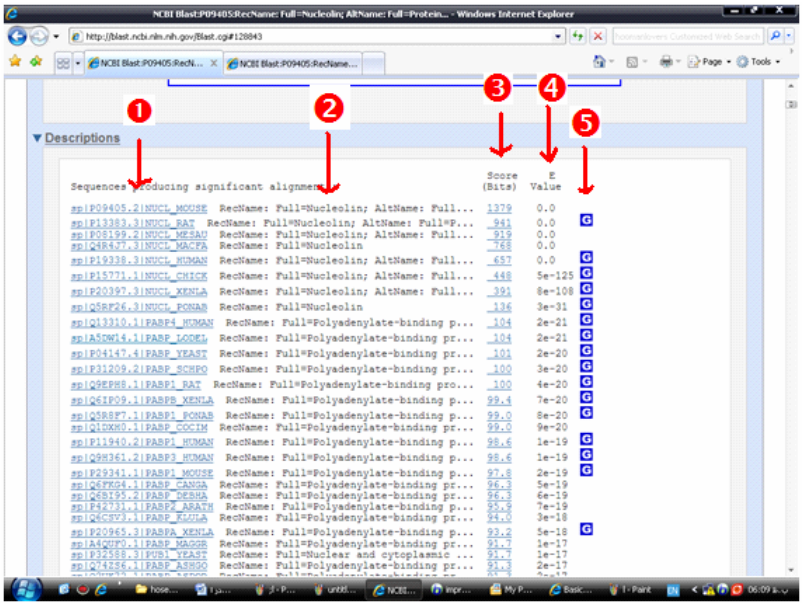
با اهمیت ترین بخش از میان تمام داده های خروجی، از نظر اکثر کاربران BLAST همین قسمت (لیست هدف) می باشد. با نگاه کردن به این قسمت می توانید بلافاصله بفهمید که آیا توالی شما با توالی های موجود در پایگاه داده مشابهت دارد یا نه، و یا اینکه آیا می توانید به این داده ها به عنوان داده هایی مشابه توجه کنید یا نه. هر ردیف موجود در این لیست، شامل ۴ ویژگی مهم می باشد:

- ✓ شماره ی مسلسل توالی در بانک داده و نام توالی: با کلیک بر روی این قسمت صفحه ای باز می شود که تمام جزییات موجود، پیرامون این توالی را به تفسیر ذکر کرده است(شکل 7- ۱۰ شماره ۱).
- ✓ Description: این قسمت در واقع بخشی از اطلاعات کامل موجود در پایگاه داده است که به شما این امکان را می دهد.

با یک نگاه و بدون صرف وقت زیاد بفهمید که آیا توالی مورد نظر می تواند جواب سوالتان باشد یا نه؟ البته به یاد داشته باشید که هیچ ضمانتی برای درست بودن نتیجه ای که از خواندن این قسمت به دست می آورید وجود ندارد. ما پیشنهاد میکنیم قبل از اینکه از به دست آوردن نتیجه، بیش از حد هیجان زده شوید جزییات کامل در مورد پروتیین

مورد نظر رابررسی کنید(با همان لیست کامل اطلاعات که با کلیک بر شماره ی مسلسل به دست می آید.)(شکل ۲-۱۰).

- ✓ امتیاز بر حسب bit: اعدادی که بر اساس یک مقیاس اندازه گیری آماری به دست آمده و به ترتیب از بزرگ به کوچک مرتب شده اند. هر چه این عدد بزرگتر باشد یعنی میزان شباهت نیز بیشتر است. توالی هایی که bit score آن ها از ۵۰ کمتر باشد ارزشی ندارند(شکل ۷-۱۰ شماره ۳).
- ✓ عدد E(مقدار مورد انتظار): به مانشان می دهد شباهت میان توالی یافته شده و توالی مورد نظر ما چه مقدار بر اساس تصادف بوده است. عدد E در واقع یک مقیاس بسیار مهم و معنادار آماری است که در اختیار شما قرار گرفته است. (کادر عدد E، شباهت ها و همولوگی ها را بخوانید) هر چه میزان عدد E پایین تر باشد بدان معنی است که توالی یافته شده شباهت بیشتری به توالی مورد سوال دارد.هر گاه عدد E پایین بود، می توانید این اطمینان را پیدا کنید که توالی یافته شده کاملا همولوگ با توالی شماست. برای مثال، توالی هایی که کاملا همسان با توالی شما هستند(همولوگ) عدد E تقریبا برابر با ۰. دارند در حالی که توالی هایی که عدد E آن ها بالاتر از ۰,۰۰۱ باشد جزء مناطق تیره – روشن محسوب می شوند(شکل ۷- ۱۰ شماره ۴).
- ✓ ورود به منبع اطلاعات ژنومی: امروزه با وجود اطلاعات کامل در مورد بسیاری از ژنوم ها، شاید شما بخواهید منطقه ی ژنی پروتیین های مورد نظرتان را ببینید. برای این کار کافی است بر روی حرف G که در برابر هر کدام از این توالی ها نوشته شده است کلیک کنید تا مستقیما به اطلاعاتی که می خواهید برسید(شکل ۷- ۱۰ شماره ۵).



شکل ۷-۱۰

هم ترازی ها (alignments)

تمام چیزهایی که در مورد عدد E و داده های آماری گفتیم را کنار بگذارید.البته منظور این نیست که همه ی چیزهای گفته شده اشتباه بوده اند، ولی بیولوژیست ها وقتی در یک موقعیت واقعی با یک سوال مواجه می شوند تنها به alignments اعتماد می

کنند. آن ها معتقدند که اطلاعاتی که از alignments به دست می آیند هیچ گاه اشتباه نخواهد بود البته به شرطی که نحوه ی تفسیر آن ها را بدانیم.

در هر alignment BLAST موارد زیر قابل مشاهده می باشد:

- ✓ نام: اولین موردی که در این بخش دیده می شود، نام توالی است. گاهی در قسمت alignments بیش از یک نام وجود دارد. چنین موردی زمانی پیش می آید که بانک اطلاعاتی، بیش از یک توالی مشابه با توالی مورد سوال را یافته باشد، به این ترتیب نام و خصوصیات همه ی آن ها را به تفکیک گزارش می کند. چنین موردی در Non-redundant databases نیز ممکن است دیده شود(شکل ۱-۱۳).
- ✓ درصد شباهت: این عدد یک جایگزین قابل لمس برای عدد E می باشد. عدد جادویی را به یاد آورید: شباهت بیش از ۲۵٪ شباهت خوشحال کننده ای بود(این عدد از تقسیم تعداد واحد های مشابه به کل تعداد واحد های مقایسه شده ،بدون در نظر گرفتن شکاف های موجود در توالی به دست می آید.شکل ۲-۱۳).قسمت positives کسری از میزان تشابه و یکسانیست(شکل ۳- ۱۳). عدد Gap میزان واحدهایی که در سنجش مورد استفاده قرار نگرفته اند را نشان می دهد(شکل ۴-۱۳).
- ✓ طول: این عدد طولی از ۲ توالی که با هم متشابهند را نشان می دهد. در واقع این عدد بهترین راه برای فهمیدن این نکته می باشد که آیا توالی مذکور می تواند جوابی برای سوال ما باشد و یا نه؟(یعنی در واقع آیا میزان شباهتی را که ما نیاز داریم، دارد؟). بعضی مواقع ممکن است عدد E توالی هایی که همانندی بسیار کوتاهی با توالی ما دارند بسیار مطلوب باشد، در حالی که این توالی برای ما چندان ارزشی نخواهد داشت(شکل ۵- ۱۳).
- ✓ توالی فوقانی: توالی مورد سوال(شکل ۶-۱۳).
- ✓ توالی تحتانی: توالی پیشنهاد شده توسط بانک اطلاعاتی(شکل ۷-۱۳).
- ✓ ردیف میان دو توالی: در بین دو توالی ردیفی وجود دارد که شامل: علامت "+" برای آمینو اسیدهای مشابه، یک "حرف الفبا" برای آمینو اسیدهای کاملا یکسان و یک "فاصله" برای آمینو اسدهای نا هماهنگ است.(شکل ۸- ۱۳ و ۱۱)

```
Query  396  TLLAKNLSFNITEDELKEVFED---AMEIRLVSQDC---KSKCIAYIEFKSEADAEKNLE  449
            +L     L  ++TE  L E+F         IR V +D     +S G AY+ + + AD E+ LE
Sbjct  56   SLYVGELDPSVTEAMLFELFSSIGQVASIR-VCRDAVTRRSLGYAYVNYNNTADGERALE  114
```

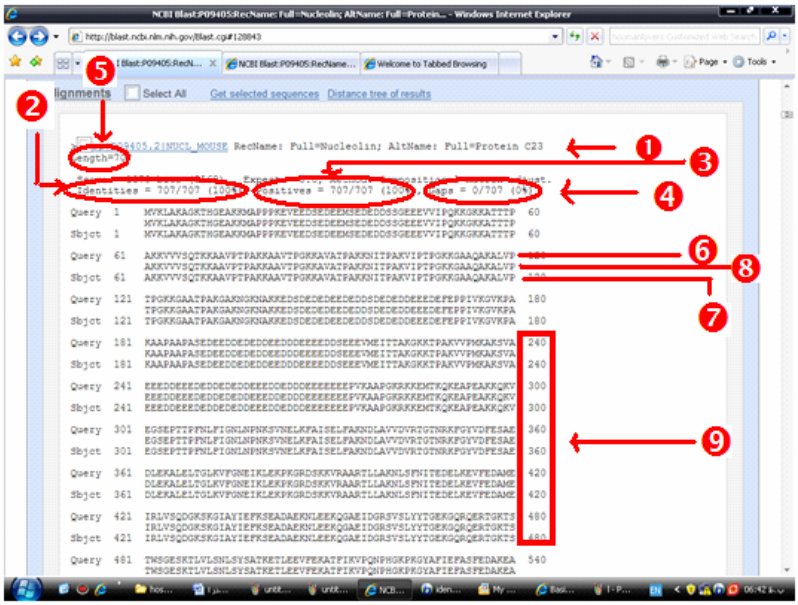
شکل ۱۱

- ✓ مناطق XXXXs : سیستم BLAST در توالی هایی که یک نوکلئوتید غالب می باشد ، خود به خود ، زنجیره ای از XXXXs را به انتهای پپتید اضافه میکند. این مناطق را مناطق " low-complexity segment" می نامند. این کار بدان دلیل انجام می شود که به علت شباهت بسیار بالای نوکئوتید های زنجیره، امکان حذف زنجیره از BLAST وجود دارد، به همین دلیل این سیستم خود به خود زنجیره ای از XXXXs اضافه می کند. به عبارتی XXXS به سیستم می گوید که شباهت مناطق را نادیده بگیرد. این عمل تنها بر روی توالی مورد سوال انجام می شود. (شکل ۱۲، به مناطق کمرنگ توجه کنید).

```
Query  1    MVKLAKAGKTHGEAKKMAPPPKeveedsedeemsededdssgeeevvIPQkkgkkatttp  60
            MVKLAKAGKTHGEAKKMAPPPKEVEE+SEDE MSEDED+   EEEVVIPQKKGKKATTTP
Sbjct  1    MVKLAKAGKTHGEAKKMAPPPKEVEEESEDENMSEDEDEKIPEEEVVIPQKKGKKATTTP  60
```

شکل 12-7

- ✓ عددها: عددهایی که در سمت راست توالی ها می بینید، مختصات منطقه ی مشابه در توالی مورد نظر شما توالی subject را معین می کنند. Balst یک تطابق منطقه ای نیز انجام می دهد که ممکن است تنها قسمتی از دو توالی را شامل شود(شکل 7- ۱۳ شماره ۹).



شکل ۱۳-۷

تعداد شکاف ها و در نتیجه ی آن ها وصله دوزی هایی که برای بستن این شکاف ها انجام می شود در یک alignment مفید و مناسب نباید زیاد باشد. حتی اگر در یک توالی نقاطی کاملا یکسان با توالی ما وجود دارد ولی بین این نقاط فاصله افتاده است نیز چندان خوشحال کننده نخواهد بود. این قاعده به خصوص در مورد توالی هایی که به مرز Twiling zone نزدیک هستند، اهمیت بیشتری دارد.

همان طور که قبلا گفته شد، گاهی مواقع در یک توالی یافته شده، بیش از یک alignment وجود دارد. این اتفاق زمانی بروز می کند که در یک توالی بیش از یک مکان مشابه با توالی مورد سوال وجود داشته باشد، که NCBI چنین توالی هایی را با یک خط باریک سیاه - میان میله ی معرف توالی مورد نظر- و EMBnet با یک خط تیره (-) به جای خط سیاه باریک نشان می دهد.

پارامترها

لازم نیست از مفهوم پارامترهایی که در پایین صفحه آمده است بترسید. اگر پیش فرض پارامترها را تغییر دهید، قسمت مربوط به این پارامترها برای شما ثبت خواهد شد.

هیچ کس نمی تواند ضمانت کند که بار دیگر شما بتوانید نتایجی مشابه با نتایج قبلی به دست آورید، حتی اگر از همان سروری که قبلا استفاده کرده اید مجددا استفاده کنید تلاش هایی که در جهت ارتقای و بهبود امکانات سرور صورت میگیرد، می تواند نتایج جستجوی شما را تحت تأثیر قرار دهد. اجزایی که ممکن است تخت روند بهبود امکانات قرار بگیرند شامل موارد زیر می شود :

- ✓ بانک اطلاعاتی
- ✓ برنامه ی BLAST

✓ پارامترهای پیش فرض سرور

در واقع، هیچ کس کنترلی بر روی این پارامترها ندارد . عمر این پارامترها با گذشت زمان تمام شده و تغییر می کنند. دست کم این پارامترهایی که در پایین صفحه آمده اند می توانند راه حلی پیرامون اینکه چه چیزی ممکن است از بین برود پیشنهاد دهد.

BLAST توالی های DNA

اگر فکر می کنید که BLAST یک رشته ی DNA مشابه با BLAST یک رشته پروتیین است، باید بگوییم که هم حق با شما هست و هم نیست. با اینکه اصول کار هر دو مشابهند، ولی استفاده از BLAST پروتیین هم سریع تر و هم دقیق تر می باشد.

اگر شما نحوه ی چارچوب خوانی رشته DNA خود را می دانید، بهتر است ابتدا خود، آن را خوانده و سپس رشته پروتیینی که از خواندن به دست آورده اید را تحت BLAST پروتیین قرار دهید، ولی اگر نحوه ی خواندن رشته ی DNA را نمیدانید، چاره ای جز استفاده از برنامه های BLAST مرتبط با نوکلئوتید ندارید.(جدول ۲ – ۷)

برنامه	سوال	بانک اطلاعاتی	کاربرد
BLASTn	DNA	DNA	توالی هایی از DNA که بسیار متشابهند.
TBLASTx	TDNA	TDNA	یافتن پروتیین مربوط به رشته DNA و ESTs
BLASTx	TDNA	Protein	تجزیه و تحلیل یک رشته ی DNA

جدول ۲-۷

T مخفف کلمه ی translation می باشد. TDNA به این معناست که شما می توانید توالی از DNA را در اختیار بانک اطلاعاتی قرار دهید تا آن را با هر ۶ چارچوب ممکن برایتان بخواند. در جاهایی که کدون پایان وجود دارد، BLAST خود به خود آن را با X جایگزین می کند. با توانایی سیستم برای خواندن هر ۶ چارچوب ممکن، دیگر لازم نیست از جا افتادن چیزی بترسید. سیستم سعی می کند هیچ امکانی را کنار نگذارد.

کدام نوع BLAST برای توالی DNA ی من بهتر است؟

هر چند جدول ۲- ۷ و کلمات مخففی که در آن نوشته شده است به نظر پیچیده می آید ولی شما نباید از آن بترسید و باید همیشه به یاد داشته باشید که « هیچ چیز» نیز بالاخره حل خواهد شد! سوالات جدول ۳- ۷ را از خود بپرسید و معین کنید که کدام برنامه برای جواب گرفتن بهتر خواهد بود.

سوال	جواب
من می خواهم در رشتهDNA های کد نشده جستجو کنم؟	بله: بهترین انتخاب BLASTn می باشد. هرگز فراموش نکنید که BLASTn تنها در DNA های با شباهت بالا جستجو خواهد کرد(بیش از ۷۰٪)

من می خواهم یک پروتیین جدید کشف کنم؟	بله: بهترین انتخاب tBLASTx می باشد.
من می خواهم پروتیین هایی که توسط رشته ی DNA ی من کد می شوند را پیدا کنم؟	بله: بهترین انتخاب BLASTx می باشد.
من می خواهم در مورد خصوصیات DNA خود مطمئن شوم؟	بله: بهترین گزینه BLASTx می باشد. البته وقتی که شما مطمئن باشید که این رشته DNA مربوط به پروتیینی خاص است، مگرنه ممکن است جز توالی های غلط باشد.

BLASTx این توانایی را دارد که خطاهای موجود در توالی شما را تصحیح کند. اگر شما از صحت توالی پروتیینی خود اطمینان ندارید، BLASTx می تواند توالی مشابهی را از یک قطعه DNA به شما معرفی کند. اگر شما به این نکته پی بردید که توالی شما دارای چارچوب ناممکنی است، بلافاصله از برنامه خارج نشوید. خطاهای مربوط به توالی ممکن است تعدادی قالب خواندن به شما پیشنهاد دهند.البته بهتر است قالب های پیشنهادی را نیز چک کنید تا از صحت آنان مطمئن شوید.

بهترین انتخاب ها برای DNA BLAST چه هستند؟

از دیدگاه سرور هیچ تفاوتی بین BLAST DNA و BLAST Protein وجود ندارد. به عبارتی طرح کلی هر کدام یکسانست به طوری که می توان با چشم بسته همان مراحل گفته شده در مورد پروتیین را در DNA نیزاجرا کرد. ولی نکات زیر را به یاد داشته باشید:

- ✓ مناسب ترین پایگاه اطلاعاتی را انتخاب کنید: مناسب ترین و مرتبط ترین پایگاه اطلاعاتی با BLAST انتخابی خود را برگزینید. مثلا اگر از BLASTx استفاده می کنید بانک اطلاعاتی باید مرتبط با DNA باشد. بسیاری از سرور ها توانایی چک کردن انتخاب شما را ندارند ،بنابراین اگر اشتباهی از طرف شما رخ دهد، بدون اینکه به شما خطایی اعلام گردد، به دنبال نتایجی خواهند بود که هرگز وجود نخواهد داشت.
- ✓ دامنه ی جستجوی خود را محدود کنید: پایگاه های ارتباطی مرتبط با DNA بسیار کند عمل می کنند، بنابراین تا جایی که امکان دارد محدوده ی جستجوی خود را محدود به نیازتان کنید. مثلا اگر اطلاعات مورد نیاز شما در مورد " Drosophila(حشره ی میوه) " است تنها در ژنوم دروسوفیلا جستجو کنید.
- ✓ در اطراف جستجو کنید: در جستجوی یک سرور مناسب و جالب برای کار خود تردید نکنید.
- ✓ از امکان "فیلتر" استفاده کنید: توالی های ژنومی مملو از تکرار هستند. با استفاده از سیستم فیلترکننده تا حد ممکن آن ها را حذف کنید.

استفاده از روش BLAST برای انجام کارها

کسانی که تجربه ی انجام تجزیه و تحلیل های بیوانفورماتیک را دارند، به این نکته دست یافته اند که تقریبا هیچ مسئله ای وجود ندارد که نتوان با BLAST حل کرد! خوشبختانه BLAST این قابلیت را دارد که برای هر چند تعداد سوال بیوانفورماتیک که درذهن داشته باشید قبل از اینکه مجبور شوید از برنامه های پیچیده ای استفاده کنید و یا حتی اینکه وقت زیادی را صرف طراحی آزمایشات

پرهزینه کنید تعدادی ایده پیشنهاد دهد.در جدول ۴-۷، ۴ مثال برای شما ذکر شده و برای هر مثال ۲ راهکار نیز طراحی شده است. یکی راهکاری جامع و پیچیده که راه حل مناسبی به حساب می آیدو دیگری استفاده از BLAST که روشی بسیار سریع و کارامد بوده و در بسیاری از موارد برای رفع نیاز کافی به نظر می رسد.(البته اگر شما وقت اضافی زیادی دارید می توانید سخت ترین راه را انتخاب کنید.این انتخاب شماست.)

نیاز	رفع نیاز با روش BLAST	رفع نیاز با روشی پیچیده
پیدا کردن ژنی خاص از میان ژن های یک ژنوم	توالی ژنوم مورد نظر را به توالی های کوچکتر همپوش تقسیم کنید(۲ تا ۵ کیلوبایت). در BLASTX از BLAST NRاستفاده می کنیم تا جستجو بدون نتایج تکراری انجام شود . اگر نام تک یاخته ی خاصی را در این قسمت وارد نکنیم، نتایج بهتر خواهد بود.	استفاده از نرم افزار Gene-prediction مقایسه با توالی های mRNAsموجود.
پیشبینی کردن در مورد فعالیت های یک پروتیین	در BLASTP از Swiss-Prot استفاده می کنیم. اگر نتایج به دست آمده در مورد تمام طول رشته خوب بود(بیش از ۲۵٪) مشکل شما حل شده است. چون عملکرد پروتیین مورد نظر شما مشابه با عملکرد پروتیین مشابه می باشد.	استفاده از: آنالیز هر قطعه و سپس تفسیر آن. آزمایشات wet-lab
پیش بینی کردن ساختمان ۳-D پروتیین	در BLASTP از)PDB پایگاه داده ای در زمینه ی ساختار پروتیین ها) استفاده می کنیم. اگر نتیجه خوب بود(بیش از ۲۵٪) می توانید بگویید که ساختار پروتیین شما شبیه به ساختار پروتیین موجود می باشد.	استفاده از مدل های همولوگ، Xrayو یا انجام آنالیزهای NMR بر روی پروتیین.
پیدا کردن اعضای خانواده ی یک پروتیین	در BLASTP از NR استفاده می کنیم.بعد از جمع آوری همه ی اعضای خانواده ی یک پروتیین می توانید یک جدولی از توالی های متعدد آماده کنید و از روی آن	با استفاده از روش کلون کردن ساخت اعضای جدید!!! استفاده از روش PCR

درختـی بـر حـسب تکامـل پـروتیین
بکشید(phylogenetic tree)

تنظیمات **BLAST**: انتخاب پارامترهای مناسب

همان طور که گفته شده است: « هر نیرو و توانایی، بدون قابلیت کنترل هیچ است». این جملـه خلاصـه وار و زیبـا بـه شـما نـشان میدهد که چگونه می توانید از BLAST بهترین استفاده را بکنید. بله درست است؛ با تنظیم کردن آن بر اساس نیازتان. در این بخـش ما به شما پارامترهای اصلی BLAST ، کارشان و چگونگی تنظیمشان را نشان می دهیم.

هنگام استفاده از BLAST اولین قانون، تعیین و روشن کردن سوال و مشکلی است که در پی پاسخ برای آن هستید. هنگـام کـار با BLAST همان طوری باشید که یک دانشمند قدیمی در انجام کارهایش بود؛ برای یافتن پاسخ هایی کـه بـه دنبـالش هـستید، ابتـدا باید بدانید که چه می خواهید؟!. جداول ۳-۷ و ۴- ۷ که سوالات احتمالی شما را به نحوی ساده و روشن بیان کرده اند می توانند در این زمینه کمکیار شما باشند.

پارامترهایی که BLAST به صورت پیش فرض استفاده می کند، ماکزیمم ممکن می باشند.اگر شما معنای هر کدام از آن هـا را ندانید، تغییر دادن آن ها، هیچ کمکی به شما نخواهد کرد. جدول ۵- ۷ تعدادی از پارامترها و دلایلی که ممکن است شما را وادار به تغییر مقادیر پیش فرض آنها کند را بیان کرده است. این استدلالات هم برای پروتیین و هم برای DNA قابل بسط هستند.

ممکن است تغییر پارامتر هر سرور کمی متفاوت با بقیه سرورها باشد. ما در این بخش بر روی NCBI متمرکز می شویم. بعـدها شـما خواهید فهمید که بسط و تطبیق این اطلاعات به بقیه ی سرورها کار راحتی است. برای انطباق جستجویتان با چیـزی کـه مـد نظرتـان هست، بهتر است همیشه از نسخه ی پیشرفته ی سرور مورد نظرتان استفاده کنید، چون معمولا امکانات بیشتری را در اختیارتـان قـرار می دهد.

دلیل و یا مشکلی که با آن مواجه هستیم	پارامتری که به تبع مشکل باید تغییر دهیم
توالی مورد نظر من، شامل تعداد زیـادی، نوکلئوتیـدهای مـشابه می باشد که می تواند باعث خطا در محاسبات شود.	Sequence filter:کـه بـه طـور خـود بـه خـود منـاطق مذکور را می پوشاند(همان مناطق XXXs)
BLASTنتیجه ای برای سوال من ارائه نمی دهد.	Substitution or gap penalties
اندازه ی عدد E توالی های مشابهی که یافته شـده در حـد مـرز می باشد.	Substitution or gap penaltiesبـرای جـستجو و یافتن مشابهت های بیشتر.
تعداد توالی های مشابه یافته شده خیلی زیاد می باشد.	-1پایگاه داده ای که استفاده می کنید، یا ۲- با تغییر کلمات کلیدی، یا ۳- تعداد توالی های مشابه را افزایش دهید، یا ۴-مقدار عدد E را افزایش دهید، یا ۵- توالی هـای بـسیار متـشابه را از محاسـبه ی خـود حـذف کنید(توالی هایی با عدد E بسیار پایین.

جدول ۵-۷

چگونه به سیستم BLAST بگوییم که کدام مناطق از توالی ها را از محاسبات خود خارج کند؟

وقتی BLAST در حال جستجو در منابع اطلاعاتی است همیشه یک پیش فرض را مد نظر دارد و آن اینکه همه ی توالی ها مشابه توالی های معمول هستند. برای مثال اگر شما در پی جستجو در یک بانک اطلاعاتی پروتیین هستید، BLAST ترکیبات پروتیین شما را مشابه با پروتیین های دیگر فرض می کند. ولی چنین فرضی همیشه مطابق با واقعیت نیست. برای فهمیدن این موضوع ادامه ی بخش را مطالعه فرمایید.

نقابدار کردن توالی های پروتیینی

در تعداد زیادی از پروتیین ها بخش هایی وجود دارند که به زبان عامیانه به هم وصله شده اند (در زبان تخصصی به این بخش ها Low- complexity می گویند).برای فهم این مطلب به مثال زیر توجه کنید. پروتیینی را در نظر بگیرید که از قطعه های پر از پرولین یا گلوتامین و یا هر اسیدآمینه ی دیگری تشکیل شده است. BLAST وقتی با چنین مناطقی روبرو می شود آن ها را مناطق هم تراز به حساب می آورد، طبیعتا عدد E مربوط به این مناطق به علت مشابهت بالا، عدد خوبی خواهد بود. متأسفانه با وجود مناسب بودن عدد E در اغلب موارد چنین مناطقی هیچ وابستگی ای به هم ندارند. در واقع این منطقه ی پر آمینواسید یک خطا برای BLAST محسوب می شود، در حالی که خود سیستم BLAST آن را نمی فهمد.

سیستم این امکان را به شما می دهد که بنا بر نیازتان از بروز چنین خطاهایی جلوگیری کنید. برای این کار وارد قسمت Algorithm parameters که در انتهای صفحه ی Blastp وجود دارد، شوید و گزینه ی Low complexity از منوی Filter and masking را در حالت انتخاب شده قرار دهید (۱-۱۴). سیستم با قرار دادن توالی از Xs به جای توالی های تکراری، خود به خود می فهمد که چنین توالی هایی را در محاسبات وارد نکند.

برای کم کردن میزان خطا، بهتر است همیشه از این فیلتر استفاده کنید.

تصور کنید پروتیینی را کلون کرده اید که دارای یک قطعه ی تکرار شونده از پرولین است(ppppppppppppppp). ولی شما بر آن بیم دارید که آیا ممکن هست خطایی در توالی پروتیین و یا حتی مشکل دیگری وجود داشته باشد یا نه. ولی اگر پروتیینی با همین خصوصیت بیابید، شما مطمئن خواهید شد که توالی پروتیینی شما نیز صحیح است.

برای پیدا کردن چنین پروتیینی، توالی PPPPPPPPPPPPPPPPP را در جعبه ی جستجو بنویسید و فراموش نکنید که گزینه ی Low complexity باید از حالت انتخاب خارج باشد. اگر پروتیینی با این خصوصیات موجود باشد، BLAST آن را برای شما خواهد یافت.

مناطقی وجود دارند که حضورشان در توالی های پروتیینی بسیار شایع می باشد مانند "Zn finges in tandems" و "Fibronectines". اگر پروتیین شما نیز حاوی چنین مناطقی باشد، به علت حضور این مناطق در بسیاری از پروتیین ها، تعداد نتایجی که به شما گزارش می شود خیلی بالا خواهد بود، که البته بسیاری از این نتایج هیچ ارتباطی با هدف شما ندارند. برای جلوگیری از چنین خطاهایی حذف چنین مناطقی از محاسبه بهترین راه حل می باشد:

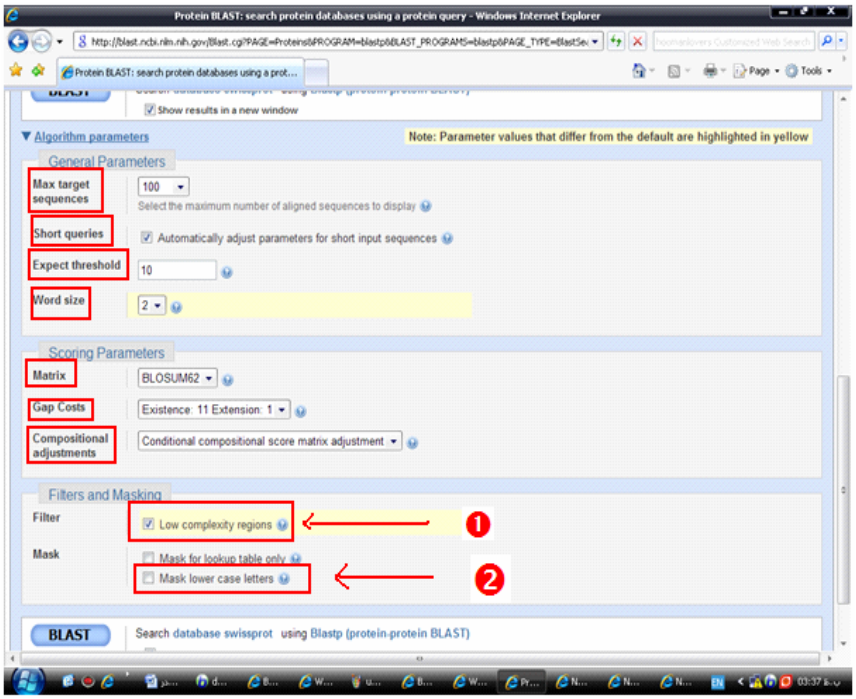
۱. از CD search, Interproscan و یا pfscan برای تشخیص این مناطق در توالی پروتیینی خود استفاده کنید.

در بخش ۶ این متدهای جستجو به تفصیل توضیح داده شد.

۲. از منابع موجود، طول قطعه ی مورد نظر را پیدا کنید.

۳. قطعات مورد نظر را با رشته ای از XXXs جایگزین کنید. سپس گزینه ی Mask lower check را در حالت انتخاب قرار دهید(شکل ۲- ۱۴).

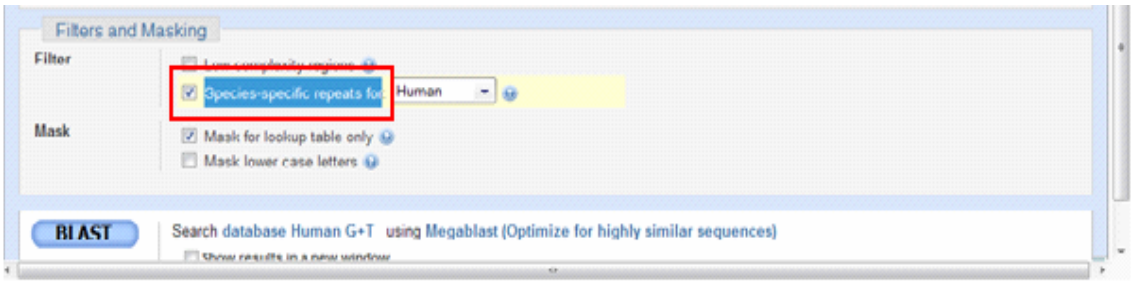
۴. بقیه ی مسیر را بر طبق توضیحات گفته شده در ابتدای بخش طی کنید.



شکل ۱۴-۷

نقابدار کردن توالی های DNA

اگر به نظر شما معنی دار کردن نتایج، در توالی های پروتیینی کار مشکلی است، بهتر است این را بدانید که در مورد توالی های DNA بسیار سخت تر می باشد. DNA مملو از توالی های تکراری است که حقیقتا هنوز معنی روشنی برای ما ندارند. بسیاری از ژنوم ها چنین قطعاتی را شامل می شوند، مخصوصا ژنوم انسانی که ۶۰٪ تکراری است. اگر شما می خواهید این توالی های تکراری از محاسبه حذف شوند، در قسمت "Species-specific repeats for" در صفحه ی BLASTn، جانور مورد نظر خود را انتخاب کرده و آن را در حالت انتخاب قرار دهید. چنین توالی هایی از نتایج شما حذف خواهند شد(شکل ۱۵-۷).



شکل ۱۵-۷

توالی های بزرگ ژنومی، خیلی هم بدون اشکال نیستند. برای مثال، در بسیاری از موارد در ژنوم یک موجود ممکن است تکه هایی از موجودی دیگر -که در واقع از ژنوم میزبان برای کلون کردن خود استفاده می کرده است- وجود داشته باشد (به همین دلیل گاهی در ژنوم انسانی، پروتیین هایی از yeast و E. coli وجود دارد!).بنابراین به یاد داشته باشید قبل از اینکه از یافتن یک ژن از ژنوم انسانی و یا هر موجود دیگری هیجان زده شوید، مطمئن شوید که آن ژن مربوط به موجود مورد نظر شما می باشد. تصور کنید چقدر مایه ی تأسف و خجالت است که بعد از ۳ سال کار بر روی « ژن مستعد کننده ی انسانی به سرطان» بفهمید که ژن مورد نظر مربوط به E.coli بوده است!

تغییر در پارامترهای مرتبط با Alignment

شکل ۱۴-۷ تمام پارامترهای قابل تغییر در سیستم BLAST را به شما نشان می دهد. پارامترهایی نظیر: عدد E، اندازه ی کلمات، انواع فیلترها و ... در بین پارامترهای موجود در این صفحه دو پارامتر مهم که با تغییر آن ها در Alignmentes تغییراتی اتفاق می افتد و نتایج متفاوت خواهد بود " Gap Costs" و " Matrix" می باشند. توالی هایی که با تغییر این پارامترها ممکن است که حذف و یا اضافه شوند توالی های نزدیک به مرز هستند(> ۰,۰۰۰۱ عدد E).

فاکتور اصلی تعیین کننده در Blast اندازه کلمات است. " کلمه" در BLAST کمترین تعداد از المان های یک توالی است که با توالی دیگر مقایسه می شود. برای مثال اگر اندازه ی کلمه را ۶ در نظر بگیریم، BLAST تمام توالی ها را(چه توالی مورد سوال و چه توالی های موجود در بانک اطلاعاتی خود) به تکه های ۶ تایی تقسیم کرده و این تکه ها را برای تعیین مشابهت ها با هم مقایسه می کند.(البته تمام حالت های خواندن ۶ تایی را در نظر می گیرد). بنابراین هر چه اندازه ی کلمه بزرگ تر باشد، عملیات سریع تر و با دقت کمتر انجام می گیرد و بالعکس.

اهمیت تغییر این پارامترها وقتی روشن تر می شود که در توالی هایی مرزی مورد استفاده قرار گیرند. به عبارتی هر گاه با تغییر "Gap Costs" و " Matrix" یک توالی مرزی حذف نشد، می توان به معنی دار بودن آن از لحاظ بیولوژیکی امید داشت.

تنظیم داده های خروجی BLAST

اگر توالی مورد سوال شما عضوی از یک خانواده ی بزرگ باشد، بانک اطلاعاتی، تعداد زیادی توالی مشابه آن را خواهد داشت و بنابراین حجم نتایج به دست آمده از BLAST بسیار زیاد خواهد بود.

گاهی این تعداد توالیهای همولوگ مانعی برای یافتن همولوگ های مورد نظر ما(همولوگ های نزدیک تر به توالی ما) خواهند بود. اگر با چنین مشکلاتی روبرو شدید، به جدول ۵-۷ مراجعه کنید. در آخرین ردیف جدول، ۵ راه حل برای حل این مشکل ارائه شده است. در این بخش روش پیاده کردن آن راه حل ها را به شما نشان خواهیم داد.

انتخاب بانک اطلاعاتی متناسب با نیازمان

بانک اطلاعاتی متناسب با نیاز خود را انتخاب کنید. اگر تعداد نتایج ارائه شده از سیستم بسیار بالا بود و شما به دنبال راه حلی برای کاستن آن هستید، از خود بپرسید که آیا بانک اطلاعاتی که انتخاب کرده بودید درست بوده است یا نه؟ مثلا می توانید به

جای NR از Swiss-Prot (که گنجایشی بسیار کمتر از NR دارد) استفاده کنید ، یا عملیات جستجو را تنها برای یک ژنوم تکرار کنید و یا اگر تنها نیاز به یافتن شباهت های ساختاری دارید، PDB را به کار برید.

محدود کردن نتایج با استفاده از Enterez query

راهی دیگر برای کاستن از تعداد نتایج استفاده از Organism box و Enterez query box (که در صفحه ی اصلی Blastp وجود دارد)می باشد(شکل۱۶-۷).

در این پنجره ها می توان از حروف ترکیبی مانند"And" و "Not" استفاده کرد. برای مثال وقتی هدف ما جستجوی توالی در گونه ی پروتئازها می باشد ولی نمی خواهیم که توالی مربوط به ویروس HIV باشد، در پنجره می نویسیم:

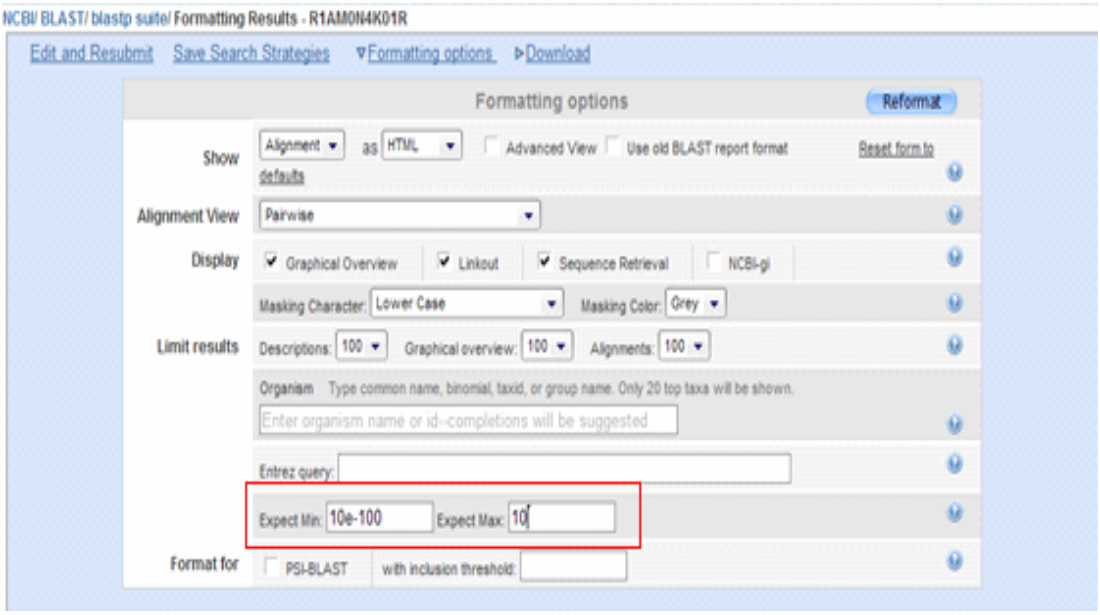
"p rotease NOT hiv1[Organism]

دقت داشته باشید که گذاردن علامت " مهم است. اگر اطلاعات بیشتری در این زمینه نیاز دارید در مقابل این جعبه ها بر روی علامت سوال کلیک کرده و گزینه ی More را انتخاب کنید.

Expect

یک راه دیگر برای کم کردن تعداد نتایج، تغییر در محدوده ی عدد E می باشد. قبلا گفته شد که هر چه میزان عدد E کوچکتر باشد بدان معنی است که توالی مذکور شباهت بیشتری به توالی مورد سوال دارد. بنابراین با تعیین یک عدد کوچک برای عدد E در واقع سیستم را وادار می کنیم که شبیه ترین توالی ها را برای ما جدا کند.

این سیستم به ما این امکان را می دهد که عدد E را از هر دو طرف محدود کنیم و به این ترتیب تعداد توالی هایی که نیمه مشابهند کمتر گزارش خواهد شد. برای مثال در شکل ۱۶ ما به سیستم دستور داده ایم که بهترین و شبیه ترین توالی ها را حذف کند. (عدد E>10e – 100)



شکل ۱۶-۷

به وسیله ی BLAST–PSI جستجوی خود را تکرار کنیم

گاهی اوقات با استفاده از BLAST به تمامی نتایج دلخواه دست نخواهیم یافت. فرض کنید شما به دنبال تمام پروتیین هایی هستید که مرتبط با توالی شما هستند. با استفاده از BLAST تنها توالی هایی را به دست می آورید که شباهت زیادی به توالی شما دارند. ولی بقیه اعضای دور خانواده چه طور؟

BLAST–PSI چنین کاری را برای شما انجام می دهد. این سیستم ابتدا به دنبال توالی های بسیار مشابه می گردد و سپس کم کم و با دقت، دایره ی جستجوی خود را وسعت می بخشد تا تمام توالی هایی را که در ابتدا جالب به نظر نمی آمدند ولی پاسخی برای خواسته ی ما بوده اند نیز پیدا شوند.

BLAST–PSI هر بار سعی تازه ای را آغاز می کند تا اعضای تازه ای از خانواده را بشناسد. در واقع این سیستم به خصوصیات شاخص همه ی توالی های یافته شده توجه می کند تا اعضای دیگر خانواده را هم شناسایی کند.

BLAST–PSI توالی های پروتیینی

سوالی که در این قسمت بر روی آن کار خواهیم کرد این است که آیا "leghemoglobin" که پروتیینی در گیاه با وظیفه ی گرفتن اکسیژن می باشد شباهتی با هموگلوبین انسانی دارد یا نه؟

۱. همان صفحه ی اصلی BLASTp را باز کنید(شکل ۳).

۲. در پایین صفحه قسمتی به نام program selecting وجود دارد. مطابق شکل ۱۷، BLAST–PSI را در حالت انتخاب قرار دهید.

شکل ۱۷–۷

۳. شماره و یا فرمت FASTA ی توالی مورد نظر را در جعبه جستجو وارد کنید.

قوانین مربوط به این بخش دقیقا همانند قوانین مربوط به پر کردن این بخش در Blastp می باشد. یعنی اگر چنین توالی قبلا در بانک اطلاعاتی موجود بوده است کد شناسایی ان را وارد می کنیم و در غیر این صورت فرمت FASTA ی آن را می نویسیم. در این جا ما کد شناسایی هموگلوبین یعنی p01922 را وارد می کنیم.

۴. از منوی Database بانک اطلاعاتی PBD را انتخاب می کنیم.

PBD بانک اطلاعاتی با پهنای کم، شامل ساختمان های سه بعدی پروتیین هاست که برای به دست آوردن نتیجه ی سوال ما مناسب می باشد.

هدف از این جستجو، این است که توالی را از پایگاه اطلاعاتی PDB پیدا کنیم که درواقع همان Leghemoglobin بوده و سپس بر اساس آن نتیجه بگیریم که بین هموگلوبین و Leghemoglobin شباهت ساختاری وجود دارد.

۵. اگر BLAST-PSI پروتیین همولوگی با پروتیین شما نیافت میتوانید این دفعه شانس خود رادر بانک اطلاعاتی NR به جای PDB امتحان کنید.(NR بانک اطلاعاتی شامل تمام توالی های پروتیینی بدون توالی های زاید می باشد). با این فرایند توالی هایی که مابین توالی شما و توالی های یافته شده در PDB هستند و ممکن است توالی مورد نظر شما مابین آن ها قرارگرفته باشد، پیدا می شوند.

۶. بر BLAST کلیک کنید تا صفحه ی حاوی نتایج باز شود. صبور باشید. چون زمان لازم برای باز شدن صفحه نهایی ممکن است خیلی بیشتر از مقداری باشد که خود برنامه نشان می دهد.

۷. تجزیه و تحلیل داده ها.

در صفحه ی نتایج، ۲ دسته توالی دیده می شوند. یک دسته که اکثریت را هم تشکیل می دهند، توالی هایی با شباهت بالا هستند و دسته ی دوم که تعدادشان کمتر است، توالی هایی با وابستگی کمتر هستند. ولی نتایج متأسفانه چندان خوشحال کننده نخواهند بود زیرا همان طوری که انتظار می رود، توالی ها اکثرا مربوط به هموگلوبین و تعدادی نیز مربوط به میوگلوبین می باشند.

به شکل شماره ی ۷-۱۸ توجه کنید: علاوه بر ستون هایی که در صفحه نتایج BLAST در قسمت لیست هدف داشتیم، در این جا ۳ ستون دیگر هم وجود دارد:

- ✓ جعبه ی نشان گذاری: اگر این گزینه در حالت فعال قرار گرفته شده باشد، یعنی از محتوای این توالی در جستجوی بعدی استفاده خواهد شد.
- ✓ نشان new: در هر بار تکرار BLAST-PSI این علامت در کنار توالی هایی قرار می گیرد که در دورهای قبلی وجود نداشته اند و این اولین باری است که گزارش می شوند.
- ✓ قرص های سبزرنگ: این علامت در کنار توالی هایی قرار می گیرد که BLAST-PSI از آن ها برای یافتن توالی های جدید استفاده کرده است.

اگر از روی توضیحات یک توالی متوجه شدید که آن توالی ارتباطی به توالی اولیه ی شما ندارد، آن را از حالت انتخاب شده خارج کنید. به این صورت توالی های نامربوط برای دور دوم تکرار، انتخاب نخواهند شد.

آستانه ی تعیین شده در مقابل گزینه ی iteration 2 BLAST-Run PSI که در پایین این بخش قرار دارد را تغییر داده و سپس آن را انتخاب کنید. به این ترتیب دستور انجام تکرار جستجو برای بار دوم را صادر کرده اید.

۸. بعد از باز شدن صفحه ی نتایج، مرحله ی ۷ را برای ۹ تا ۱۰ بار تکرار کنید.

در هنگام استفاده از BLAST-PSI چگونه از بروز خطا جلوگیری کنیم؟

در مورد مثال مطرح شده در این بخش همان طور که دیدید، دستور تکرار را چندین بار صادر کردیم. در واقع تا جایی که توالی های جدید ثبت شده در صفحه ی نتایج مرتبط با هموگلوبین باشند، این کار را تکرار می کنیم.

به عبارتی دیگر، گاهی پیش می آید که مسیر تکرار، خود به خود قطع می شود. مثلا در مورد مثال ذکر شده، اگر در دور دوم تکرار، تمام توالی های جدید، الکل دهیدروژناز بودند، ادامه ی تکرار کار عاقلانه ای نخواهد بود. متأسفانه برای تعیین اینکه چه زمانی باید عملیات جستجو را قطع کرد و یا چه زمانی هنوز باید ادامه داد، قانون مشخصی و ساده ای وجود ندارد. هر رخداد تازه ای تنها ۲ معنی می تواند داشته باشد: یا ادامه ی کار باید متوقف شود و یا چیز تازه ای پیدا کرده اید.

به یاد داشته باشید که در بین دو تکرار، نه تنها می توانید، بلکه موظفید که توالی های بی ارتباط را از حالت انتخاب خارج کنید.

چگونه توالی های صحیح را انتخاب کنیم؟

تعیین توالی های مناسب برای نگهداری آن ها و حذف توالی ها نامناسب اصلی ترین مشکلی است که در BLAST-PSI ممکن است با آن مواجه شویم. سیستم BLAST-PSI خود، به صورت اتوماتیک بر حسب آستانه ای که به آن داده ایم این انتخاب را برای ما انجام می دهد. البته ما خود به صورت دستی نیز می توانیم با انتخاب کردن و یا نکردن توالی ها به صورت مجزا این کار را انجام دهیم.

تعیین آستانه در عین حال هم می تواند مفید باشد و هم مضر:

- ✓ اگر آستانه را بسیار پایین تعیین کنیم، هیچ توالی یافت نخواهد شد.
- ✓ اگر آستانه را خیلی بالا تعیین کنیم، تعداد زیادی توالی نامربوط نیز در یافته های ما قرار خواهند گرفت.

قرار گیری توالی های نامربوط در نتایج را می توان به حمله ی کنه تشبیه کرد: مدام بدتر و بدتر می شود. اولین توالی نامربوطی که در نتایج قراربگیرد، باعث جذب توالی های نادرست دیگری می شود.

تنها علامتی که نشانه ی آستانه ی پایین است. یافت نشدن موردی است. بهترین راه، استفاده از اعداد پیش فرض و ارزیابی آن هاست. بعد از هر دوره تکرار، برای نتیجه ی مناسب تر، بهتر است که پس از بررسی نتایج به دست آمده، توالی های نامرتبط را حذف و توالی های مرتبط را اضافه کنیم. این کار هر چند که می تواند وقت گیر باشد ولی بیشتر اوقات نتیجه خوبی در بر دارد.

راه دیگر عوض کردن توالی مبدأ است. یعنی توالی هایی را که توجه ما را به خود جلب کرده اند را به عنوان توالی مبدأ قرار دهیم. به این ترتیب می توانیم انتظار داشته باشیم که در صورت مناسب بودن آن توالی، توالی اولیه ی خودمان به عنوان یکی از توالی های مشابه یافت شود.

وقتی محدوده ای از یک توالی، خود به تنهایی می تواند شبیه به یک پروتیین مستقل باشد بنابراین وجود پروتیین هایی با منطق چندگانه عجیب نخواهد بود. ممکن است مناطقی مشابه در پروتیین هایی کاملا متفاوت دیده شوند.

اگر پروتیین مورد نظر شما یک پروتیین چندگانه است BLAST-PSI را به اشتباه خواهد انداخت. یعنی هر مکنطقه می تواند به عنوان یک منطقه برای جستجو در نظر گرفته شود و در نتیجه ی آن تعداد زیادی پروتیین نامرتبط یافت شود.تجزیه و تحلیل نتایجی که از چنین پروتیین هایی به دست می آید کار دشواری خواهد بود.

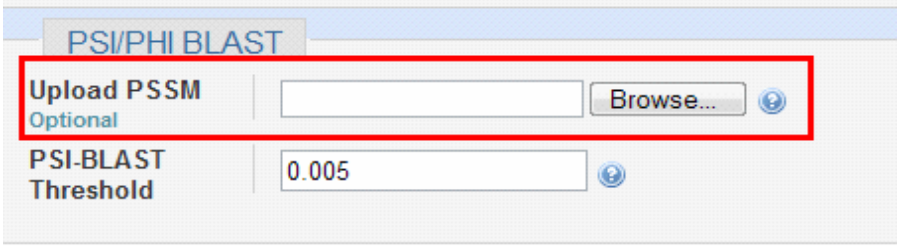
وقتی با چنین ابهاماتی روبرو شدید بهترین راه تکه تکه کردن توالی به مناطق موجود در آن است. بعد از جداسازی مناطق موجود در توالی می توان هر کدام از آن ها را به صورت یک توالی مجزا در نظر گرفت. برای شناساسیی مناطق موجود در توالی، می توان از NCBI CD و یا EBMnet pfscan استفاده کرد(برای کسب اطلاعات بیشتر به فصل ۶ مراجعه کنید).

اگر توالی شما بیش از ۲۰۰ آمینو اسید داشت ولی منطقه ی خاصی را در آن نیافتید، به عنوان یک قاعده همیشه در ذهن داشته باشید که در چنین مواقعی، توالی خود را به بخش های کوچکتر(هر ۲۰۰ اسید امینه) تقسیم کنید و هر کدام را به صورت مجزا مورد تجزیه و تحلیل قرار دهید. اشکالی که ممکن است براین مسیر وارد باشد این است که کنار هم قرار دادن نتایج حاصل، مانند یک پازل کار خسته کننده ای است. اما گاهی این کار نیز لذت بخش خواهد بود.

چگونه مناطق موجود در پروتیین خود را با استفاده از BLAST و PSI–BLAST شناسایی و جدا کنیم؟

در فصل ۶ چگونگی یافتن مناطق شناخته شده در توالی پروتیین را با استفاده از سرویس CD (BLAST–Rreverse-PSL یا BLAST–rps) را یاد گرفتید. حتی برای این کار می توان از BLAST برای یافتن این مناطق و استفاده در جستجو استفاده کرد. یک "منطقه"در BLAST با نام "PSSM"(Position Specific Substitution Matrix) شناخته می شود.

هنگامی که از BLAST–PSL استفاده می کنید، می توانید داده های خود را به صورت PSSM دریافت کنید. برای این کار می توانید از قسمت Upload PSSM ازقسمت Algorithm parameters استفاده کنید(شکل ۱۸–۷).



شکل ۱۸–۷

با استفاده از این سیستم، سرور PSSM مورد نظر شما را جدا کرده و در پایگاه داده ها به دنبال آن خواهد گشت.

جستجوی توالی های مشابه با سرویسهای مجانی در اینترنت

دستورالعمل ها و راهکارهای استفاده از BLAST بسیار زیاد است. برای کسب اطلاعات بیشتر در این زمینه می توانید از قسمت Help موجود در بخش BLAST NCBI استفاده کنید.(در بالای صفحه مزکور گزینه ی Help را انتخاب کنید.)

اگر از ابتدای بخش با کتاب برای یافتن پاسخ سوال مطرح شده پیش آمده باشید حتما متوجه شده اید، گاهی پیش می آید که به علت هجوم استفاده کنندگان، دستیابی به نتایج بسیار به آرامی و کند صورت می گیرد. ولی خوشبختانه برای این مشکل راه حل آسانی وجود دارد. سرویس های مجانی ارائه دهنده ی سرویس BLAST در اینترنت زیاد هستند. جدول ۶– ۷ شامل تعدادی از قوی ترین و به روز ترین سرویس های مذکور می باشد.

قاره و یا کشور پشتیبان	برنامه ارائه داده شده	آدرس سایت
آمریکا	BLAST/PSI-BLAST	http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi
اروپا	BLAST	/http://www.expasy.org/tools/BLAST

اروپا	BLAST	http://www.ch.embnet.org/software/bBLAST.html
اروپا	BLAST	http://www.ebi.ac.uk/Tools/BLAST2/
ژاپن	BLAST/PSI-BLAST	http://BLAST.ddbj.nig.ac.jp/top-e.html

جدول ۷-۶

WU-balst سایتی مشابه با BLAST-NCBI بوده که متعلق به دانشگاه واشنگتون می باشد. شاید تعجب کنید اگر بگوییم BLAST-WU در درج نقایص بسیار حساس تر و توانمندتر از BLAST-NCBI است. بیان دلایل و استدلالات این ادعا ساعت ها زمان می برد. شاید بهترین استدلال این است که پیشنهاد کنیم خود این سایت را امتحان کنید تا برایتان اثبات شود.جدول ۷ راهنمایی جهت استفاده از این سایت می باشد.

آدرس	شرح و توضیح
blast.wustl.edu/	The Home of WU-BLAST (no online server)
tigrblast.tigr.org/tgi/	WU-BLAST at TIGR, with a twist on bacterial genomes and nucleotide analysis
www.genome.wustl.edu/tools/blast/	blastn and tblastn, for digging into complete genomes
www.ebi.ac.uk/blast/	All flavors of BLAST and WU-BLAST
	Another complete WU-BLAST server

هرچند BLAST رایجترین راه برای جستجو در منابع اطلاعاتیست ولی تنها راه ممکن نیست. سه راه جایگزین که می توان برای BLAST پیشنهاد داد شامل:

- ✓ Smith and Waterman(SSEARCH): هر چند کندتر از BLAST می باشد ولی دقت آن بیشتر است.
- ✓ FASTA: کمی از BLAST کندتر ولی در سنجش DNA دقیق تر می باشد.
- ✓ BLAT: توانایی سریع تعیین محل cDNA در ژنوم و همچنین یافتن پروتیین های مشابه در یک ژنوم (پستاندار در مقایسه با پستاندار) را دارد.

عملا نحوه ی استفاده از این راهکارها بسیار شبیه به BLAST بوده و تطبیق آموزه های این بخش با هر پایگاه اطلاعاتی مشابه کار چندان سختی نیست. در جدول ۸ تعدادی از پایگاه های ارائه دهنده ی این خدمات به شما معرفی می شوند.

هیچ کدام از این برنامه ها امکانی برای حذف مناطق Low-complex را ندارند.

قاره و یا کشور پشتیبان	برنامه ارائه داده شده	آدرس سایت
آمریکا	FASTA	http://fasta.bioch.virginia.edu/fasta
اروپا	FASTA	http://www.ebi.ac.uk/fasta33/

اروپا	SSEARCH	http://www.ch.embnet.org/software/GMFDF_form.html
ژاپن	SSEARCH/FASTA	http://BLAST.ddbj.nig.ac.jp/top-e.html
آمریکا	BLAT	http://genome.ucsc.edu/

فصل هشتم

مقایسه کردن دو توالی

در این فصل می خوانیم:

- ✓ ساختن یک نقشه dot plot از طریق اینترنت
- ✓ تشخیص تکرارها در یک توالی پروتئینی
- ✓ فهمیدن تطابقهای موضعی و کلی (local and global alignments)
- ✓ پیدا کردن بهترین تطابقهای موضعی بین دو توالی مورد نظرتان

.....

"ارزش واقعی شما کاملابستگی به همان چیزی دارد که شما با ان مقایسه می شوید."

— Bob Wells

در این فصل، ما به شما روش هایی را نشا ن می دهیم که برای مقایسه کردن دو توالی پروتئینی- یا دو توالی DNA - در دسترس هستند. بزرگان بیوانفورماتیک در زبان ویژه خودشان به این آنالیز ها " مقایسه های جفتی (Pairwise) " می گویند. به این دلیل که آنها با یک جفت از توالی ها مرتبط هستند. اینها بسیار مفید هستند و میتوانند به شما کمک کنند تا:

- ✓ خود را متقاعد کنید که دوتوالی در واقع همولوگ هستند
- ✓ دریابید آیا توالی شما دارای یک دامین مشترک با توالی دیگر می باشد یاخیر
- ✓ محل دقیق ویژگیهای مشترک، نظیر پلهای دی سولفیدی و جایگاههای فعال را مشخص کنید
- ✓ یک ژن و محصول آنرا مقایسه کنید.

معمولا زیست شناسان از مقایسه های جفتی استفاده می کنند تا نتایج جستجو در پایگاههای اطلاعاتی را تصفیه نمایند و آنالیز های مشروحی را انجام دهند. این موضوع یک امر متداول است که بعد از جستجو در بانک های اطلاعاتی از تکنیک های جفتی استفاده شود تا کنترل گردد که آیا بعضی از تطابقهای گزارش شده واقعا جالب توجه هستند، یا تنها یک جرقه موردی بوده اند.

در این فصل، ما به شما نشان خواهیم داد که کدامیک از روشها، نیازهای شما را بیشتر رفع خواهند کرد. همینطور به شما نشان خواهیم داد که چگونه با ستفاده از اینترنت روشی را در مورد توالی مورد علاقه خود به انجام رسانید.

مطمئن شوید که توالی ها و روشهای صحیح را در دست دارید

هنگامیکه یک آنالیز جفتی را انجام می دهید، ظریفترین تصمیمی که می گیرید، انتخاب دو توالی است که می خواهید آنها را مقایسه کنید. انتخاب این دو توالی تا حدودی شبیه انتخاب دو نفری است که می خواهید برای مسابقه مشت زنی آنها را برگزینید. هدف آن است که مهیج ترین نبرد را تدارک ببینید.

از مقایسه جفتی به منظور یافتن توالی همولوگ توالی مورد نظر خود استفاده نکنید. اینکار زمان زیادی می گیرد. اگر می خواهید توالی مورد نظرتان را با تک تک توالی های دیگر در پایگاه اطلاعاتی مقایسه نمایید، از یک برنامه جستجوی بانک های اطلاعاتی، نظیر BLAST استفاده کنید (برای اطلاع بیشتر از BLAST به فصل ۷ مراجعه فرمایید).

اگر فصل ۷ را مطالعه کرده اید، شما در موقعیتی قرار دارید که ادعا نمایید برنامه های جستجوی بانک های اطلاعاتی تنها به مقایسه جفتی بین یک توالی مورد نظر و تمام توالی های موجود در آن بانک اطلاعاتی می پردازند. بنابراین، نیازی نیست که مقایسه جفتی دیگری انجام دهید. اما ایا واقعا همینطور است؟ یک لحظه صبر کنید. گرچه این موضوع صحیح است که برنامه هایی نظیر BLAST بانک های اطلاعاتی را از طریق مقایسه های جفتی جستجو می کنند، این برنامه ها برای سرعت، و نه برای دقت تطابق بهینه سازی شده اند. برنامه هایی که ما در اینجا به آنها می پردازیم درست برعکس این حالت عمل می نمایند. آنها چنان بهینه سازی شده اند تا دقیقترین نتایج را ارائه دهند. یعنی همان چیزی که در مورد توالی های به دقت انتخاب شده مد نظر شماست.

انتخاب توالی های صحیح

یک دلیل خوب برای انجام مقایسه جفتی بین دو توالی، وجود شکی قوی برای همولوگ بودن دو توالی است. یعنی آنها یک جد مشترک دارند (برای توضیح کامل به فصل ۷ مراجعه فرمایید). چنین توالی هایی معمولا دارای ساختمان ۳ بعدی و اعمالی شبیه هم هستند.

بهترین راه برای یافتن یک توالی که با توالی مورد نظر شما همولوگ است، جستجوی بانک اطلاعاتی با برنامه BLAST می باشد. توالی های خود را براساس معیارهای (محافظه کارانه) زیر انتخاب کنید:

- ✓ **توالی DNA** - حداقل ۷۰ درصد مشابهت دربیش از ۱۰۰ باز بین توالی مورد نظر (Query) و حاصل جستجو (Hit)، یا E-value کمتر از ۴-۱۰ (برای اطلاعات بیشتر در مورد E-value به فصل ۷ مراجعه کنید).
- ✓ **توالی پروتئینی**– بیش از ۲۵ درصد مشابهت در ۱۰۰ اسید آمینه بین توالی مورد نظر (Query) و حاصل جستجو (Hit) یا E-value کمتر از $۱۰^{-۴}$.

متاسفانه، این معیارها تنها یک مورد استفاده از مقایسه است، و یک گارانتی به حساب نمی آید. اگر حاصل جستجوی شما نمره ای نزدیک آستانه دارد، ممکن است با توالی مورد نظر شما همولوگ نباشد، یا ممکن است چنان ارتباط دوری داشته باشند که تطبیق دادن صحیح آنها مشکل باشد. این همان جایی است که روشهای مقایسه جفتی وارد صحنه می شود. این روشها به شما کمک می کنند تا تصمیم بگیرید که حاصل جستجوی شما تا چه حد بصورت معنی داری با توالی مورد نظر شما تطابق دارد.

انجام یک جستجو در بانک های اطلاعاتی ، شبیه جستجوی سریع آگهی های طبقه بندی شده برای یافتن یک آپارتمان جدید است. شما ممکن است موارد جالبی پیدا کنید، اما نمی توانید مطمئن باشید که محل زندگی کاملا (یا شاید به اندازه

کافی) مناسبی پیدا کرده اید، مگر آنکه آپارتمان را از نزدیک ببینید. آن آپارتمانی که ادعا شده پر پرستیژ و در بالاترین طبقه (Penthouse) مجهز به سیستم تهویه مطبوع و منظره خیره کننده است ، ممکن است در حقیقت تنها یک اتاقک قدیمی دارای سوراخهای فراوان در دیوارهایش بر بالای ساختمان یک کارخانه باشد.

انتخاب یک توالی از خروجی جستجو در یک بانک های اطلاعاتی، در حقیقت ایجاد تلفیق مناسبی از اطلاعات جدید و قطعیت است.. شما خواهان یک توالی هستید که به اندازه کافی شبیه به توالی مورد نظرتان (Query) باشد تا بتوانید مطمئن باشید که این همان همولوگ است. از طرف دیگر ، شما مایلید این توالی از نقطه نظر بیولوژیکی جالب باشد . تا این مقایسه ها اطلاعات جدیدی را در اختیار شما قراردهند . بطور خلاصه ، هدف شما یافتن یک توالی Swiss-Prot است که وسیعاً مورد استناد قرار گرفته باشد و با query شما مطابقت داشته باشدو E-value آن کمتر از $۱۰^{-۴}$ باشد. در واقع ما می خواهیم که به تلفیق مناسبی از اطلاعات جدید و قطعیت برسیم.

هیچ قانون نمی گوید شما باید توالیتان را از جستجو در بانک اطلاعاتی بدست آورید، اما این راه خیلی مطمئن است. راه دیگر این است که دو توالیتان را بوسیله یک محک آزمایشی بدست آورید، یعنی دو توالی را انتخاب کنید که عملکرد مشابه یا ویژگی مشابه دیگری دارند. برای این هدف ، شما ممکن است از سیستم بازیابی توالی[1] (SRS) که امکان شناسایی توالی ها را با استفاده از کلید واژه برای شما فراهم می کند(برای اطلاعات بیشتر فصل ۲ را ببینید).

اگر شما فقط یک توالی دارید، می توانید (و توصیه ما نیز همین است که) با همان یک توالی یک مقایسه جفتی انجام دهید. فقط کافی است وانمود کنید توالی دوم کاملاً شبیه به اولی است. این ممکن است کمی غیرمنطقی باشد، اما راه خوبی است برای آنکه بدنبال ویژگی های جالبی در توالیتان بگردید، مثل :

- ✓ دامین های تکراری
- ✓ نواحی با یک موتیف[2] کوچک که چندین بار تکرار شده (نواحی با پیچیدگی کم)
- ✓ نواحی پالیندروم[3] (قسمتهایی از DNA که از دو طرف به یک شکل خوانده می شوند) و ساختمانهای ثانویه بالقوه در RNA

تکرارهای داخلی، می توانند به روشن کردن فعالیت پروتئین شما کمک کنند. برای مثال، با کشف کردن یک ناحیه تکرار شده در ژن مستعد کننده به سرطان سینه از تیپ یک (Brca 1) می توانید یک پروفایل دامنه ای برای این دامنه بسازیدو نقش Brca 1 را در مکانیسم های ترمیم DNA را مشخص کنید.

انتخاب متد صحیح

مقایسه کردن توالی ها یکی از مشکل ترین کارهایی است که بیولوژیست ها با آن روبرو هستند. در واقع، هنوز کاملاً مشکل مقایسه کردن توالی ها برطرف نشده است، به همین دلیل است که بیش از یک راه حل برای این کار وجود دارد. جدول ۱-۸ سه روش را لیست می کند و نشان می دهد هر کدام در چه موقعیتی بهتر است به کار بروند.

همیشه شروع کردن با dot plot استراتژی خوبی است (ما در قسمت بعد ، به جزئیات dot plot خواهیم پرداخت). dot plot ابزار بسیار قدرتمندی است که به شما یک تصویر سریع و کلی میدهد. اگر کمی با تجربه باشید، با یک نگاه ساده به یک dot plot فوراً متوجه خواهید شد که چه چیزی قرار است اتفاق بیفتد. بنابر این dot plot یک ابزار ایده آل برای تصمیم گیری راجع به قدم بعدی آنالیز شماست.

---

[1] Sequence Retrieval System
[2] motif
[3] palindromes

Method Name	*Situation*
Dot plot	General exploration of your sequence: Discovering repeats Finding long insertions/deletions Extracting portions of sequences to make a multiple alignment
Local alignments	Comparing sequences with partial homology: Making high-quality alignments Making residue-per-residue analysis
Global alignments	Comparing two sequences over their entire length: Identifying long insertions/deletions Checking the quality of your data Identifying every mutation in your sequences

جدول ۱-۸: انواع مختلف مقایسه های جفتی

با این حال dot plot به تنهایی برای یک آزمایش کاملا طراحی شده کافی نیست. dot plot ها، alignment های واقعی نمی سازند، بلکه فقط نشانه ای کلی میدهند. آنها به وضوح نمیگویند چگونه آسیدامینه یا نوکلئوئیدهای دو توالی با یکدیگر مرتبط هستند. برای رسیدن به پاسخ این سؤال باید یک alignment بسازید. دو نوع alignment وجود دارد: alignment های کلی (هنگامیکه توالی در تمام طول خود هم تراز یا ردیف شده باشد) و alignment های موضعی (جایی که تنها در قسمت های مشابه تر دو توالی هم تراز یا ردیف شده باشند).

ردیفهای کلی معمولا برای تفسیر کردن راحتتر از انواع موضعی هستند. با این حال، زمانی که ارتباط توالی ها با بررسی تمام طول آن ها مشخص می شود، یا زمانی که توالی ها شامل جهش های جانشینی[1] یا حذف خیلی طولانی نباشند، تنها ساختن یک alignment کلی مفهوم خواهد داشت. اگر شما نمیدانید alignment های موضعی یا کلی را انتخاب کنید، همیشه نوع موضعی را انتخاب کنید. هنگامیکه شما در حال مقایسه دو توالی هستید، شروع به ساختن یک dot plot از هر توالی، در مقابل خودش، بکنید. این رویکرد تشخیص تکرارهای بالقوه درون هر توالی را آسانتر میکند. با داشتن این اطلاعات، تفسیر کردن dot plot ادو توالی آسانتر خواهد بود.

مشکلات اصلی سر راه استفاده از این روش ها انتخاب درست پارامتر های تنظیمی است، که در جدول ۱-۸ لیست شده اند. متأسفانه، در این مورد یک قانون مسلم و ثابت وجود ندارد؛ انجام یک مقایسه دقیق اغلب یک پروسه آزمایش و خطاست. در ادامه این فصل، به شما نشان خواهیم داد چگونه با تغییر این پارامتر ها به یک نتیجه قانع کننده برسید.

ایجاد یک Dot Plot

dot plot، ساده ترین وسیله برای مقایسه دو توالی است. در واقع dot plot (که با استفاده از یک کاغذ و خودکار ایجاد می شود) تنها روش مقایسه توالی است که زمانی که دسترسی به کامپیوتر ندارید، قابل استفاده می باشد. (به قسمت"راهنمای DIY برای dotplot" نگاه کنید). تکنیک dot plot ارتباط نزدیکی با متد پنجره متحرک (siding window) دارد. در متد پنجره متحرک یک window قسمتی از توالی شماست، که شما آنرا با قسمت های دیگر توالی که با آن هم اندازه هستند، مقایسه می کنید (برای بیشتر دانستن در مورد اسلاید کردن پنجره ها، فصل ۶ را ببینید).

[1] insertion

dot plot دو مزیت بزرگ دارد: آنها ساده اند و به هیچ فرضیه ی بیولوزیکی احتیاج ندارد. در نتیجه احتمال اشتباه کردن در ساخت dot plot کم است. اگر مطمئن نیستید که dot plot را کاملاً فهمیده باشید، نگاهی اجمالی به که در اقسمت"راهمای DIY برای dotplot" بیندازید.

اساس تفسیر dot plot پیشرفته ترین وسیله ی جداسازی است که در دسترس شماست، همان چشمهای شما. هنگامی که مجبور میشوید تا الگوهایی مثل خطوط مورب را مشخص کنید، هیچ چیز بر چشمهای انسان نمی تواند غلبه کند، حتی اگر سیگنال ها نویز دار[1] باشند. هر گاه شما یک سیگنال خوب را در یک dot plot می بینید ،می توانید بر سر زندگیتان شرط ببنیدید که یک اتفاق جالب در حال رخ دادن است.

راهنمای DIY برای dotplot

هنگامی که یک dot plot می سازید، مثل این است که توالی های شما با هم بازی tic-tac-toe را انجام می دهند. برای مثال، فرض میگیریم که شما می خواهید این دو توالی را با هم مقایسه کنید:

توالی ۱: THEFATCAT

توالی ۲: THEFASTCAT

تفکیک اینها از مشکل است، زیرا تقریباً شبیه به هم هستند و فقط توالی دوم یک حرف S اضافه دارد. البته این مشکل هنگام کار کردن با توالی های واقعی سخت تر خواهد شد.

برای ایجاد dot plot کاغذی بردارید، آنرا شطرنجی کنید و در بالای آن توالی ۱ را بنویسید بطوریکه هر حرف در یک ستون قرار گیرد. سپس توالی دوم را در سمت چپ کاغذ بنویسید بطوریکه هر حرف در یک خط قرار گیرد، تا هر دو یک ماتریکس بسازند.

اینجا بازی شروع میشود. هر مربع را یکی یکی ارزیابی کنید و اگر نشانه ای که در بالا و کنار (سمت چپ) قرار دارد مشابه هستند، یک علامت ضربدر در آن بزنید. مانند بازی tic-tac-toe، بلندترین خط مورب برنده خواهد بود. اما خطوط مورب برای ما قطعات همانند بین دو توالی را نشان می دهند. ماتریکس زیر به سادگی نشان می دهد که دو توالی مثال ما در تمام طول خود همولوگ هستند، به استثنای یک حرف اضافه در خط عمودی.

```
..T H E F A T C A T
T X         X       X
H   X
E     X
F       X
A         X     X
S
T X           X     X
C               X
A         X       X
T X           X     X
```

[1] noisy

انتخاب نوع درست dot-plot

در این قسمت ما ابزار بسیار قدرتمندی را بنام Dotlet را معرفی می کنیم. Dotlet ابزار بسیار خوبی است، چون استفاده از آن آسان است، رایگان است، احتیاجی به نصب کردن ندارد و تقریباً در هر کامپیوتری که به اینترنت وصل شود کار می کند.

Dot let برای توالی های با طول کمتر از ۱۰۰۰۰ اسید امینه یا نوکلئوئید ایده آل است. بنابرین برای بسیاری از پروتئین ها می تواند مفید باشد اما در مورد توالی های DNA استفاده از آن فقط محدود به توالی های کوچک می شود. اگر شما بخواهید بصورت آنلاین نگاهی به توالی های بلندتر بیندازید، باید از Dnadot (ابزاری که برای انجام dot plot سریع بین توالی های بلند طراحی شده است) استفاده کنید (جدول۲-۸).

اگر توالی شما بیشتر از ۱۰۰۰۰۰ کاراکتر دارد، هیچ ابزار مناسبی بصورت آنلاین پیدا نخواهید کرد و باید نوع قدرتمندتری از Dot let را نصب کنید. دو نوع از این برنامه ها بصورت رایگان در دسترس هستند: Dotter و Dottup. شما می توانید جزئیات مربوط به آنها و آدرس آن ها را در جدول ۲-۸ ببینید.

Name	*Used For*	*Range*	*URL*	*Platforms*
Dotlet	Proteins, DNA	10,000	`www.ch.embnet.org`	All (Java)
Dnadot	Proteins, DNA	100,000	`arbl.cvmbs.colostate.edu/molkit/dnadot/`	All (Java)
Dotter	Proteins, DNA	100,000	`www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html`	Unix, Linux, Windows
Dottup	Complete genomes, DNA	>100,000	`www.emboss.org`	Unix, Linux

جدول ۲-۸: مزایای برنامه های مختلف Dot-Plot

استفاده از **Dotlet** در اینترنت

Dotlet یکی از کاربردی ترین برنامه های dot-plot است که در اینترنت در دسترس می باشد. این برنامه Java applet است و در سرور EMBnet قابل دسترسی است. applet برنامه کوچکی است که اتوماتیک وار بر internet browser شما عمل می کند و بنابراین در کامپیوتر شخصی شما کار می کند. مزیت آن این است که تا وقتی که Java فعال است احتیاجی به نصب کردن هیچ چیز ندارید (امروزه internet browser ها بطور پیش فرض Java را دارند).

Dotlet برنامه پیچیده ای نیست، اما ویژگی های خیلی زیادی دارد. اگر شما از این ویژگی ها بدستی استفاده نکنید، ممکن است به نتیجه ی دلخواه خیلی نزدیک شوید، اما آنرا از دست بدهید. قاعده ی عمومی در استفاده از Dotlet این است که شما می خواهید سیگنال واضحی را پیدا کنید و آنقدر باید پارامتر های مختلف را تغییر دهید تا به نتیجه مطلوب برسید.

در ادامه لیست مراحل کار، ما به شما مثالهایی را از "سیگنال واضح" را نشان خواهیم داد و در مورد اینکه چگونه بهتر به آنها دست پیدا کنیم چند رهنمود خواهیم داد.

دانلود کردن Dotlet

می توانید Dotlet را براحتی با یکبار کلیک کردن، یعنی با وارد شدن به آدرس اینترنتی www.isrec.isb-sib.ch/java/dotlet/Dotlet.html دانلود کنید. هنگامی که این صفحه را باز کردید، مرورگر اینترنت شما برنامه Dotlet را دانلود خواهد کرد. اگر سرعت اینترنت شما پایین باشد، اینکار چند دقیقه طول خواهد کشید.

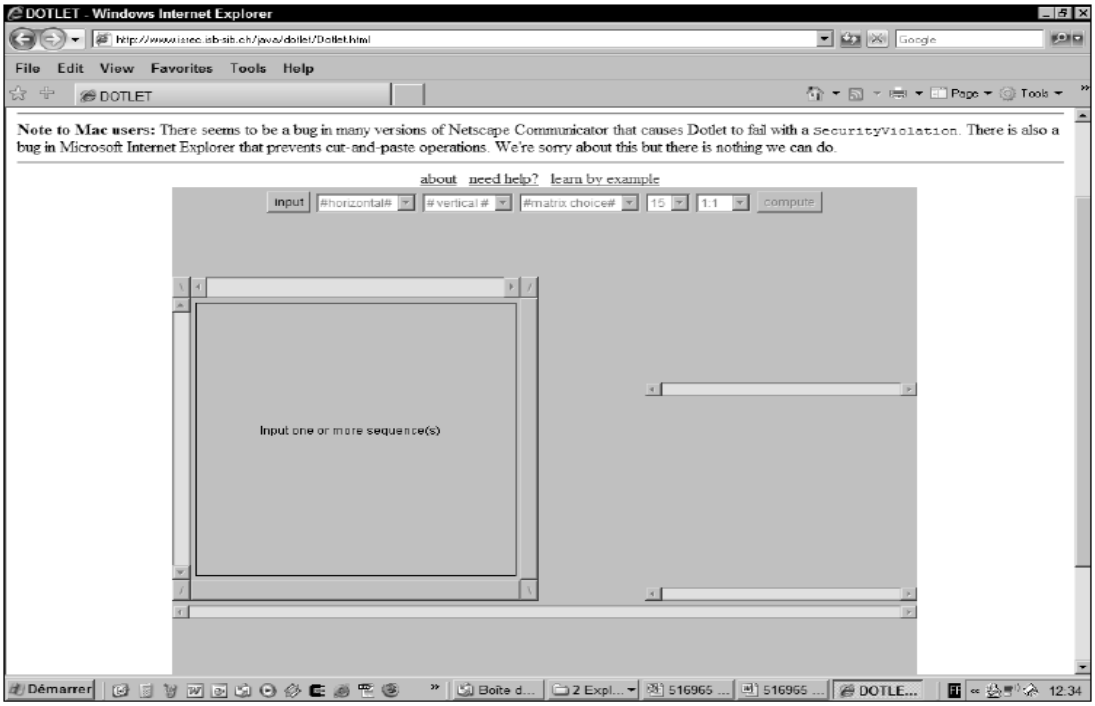
هنگامی که دانلودتان تمام شد ، مرورگر اینترنت شما صفحه ای مشابه آنچه که در شکل ۱-۸ است را نشان خواهد داد.

اگر مرورگر اینترنت شما این صفحه را بدرستی نشان نداد، به این معنی است که نمی تواند Java را راه اندازی کند. امکان دارد احتیاج داشته باشید تا آخرین نسخه از internet explorer یا Netscape را نصب کنید تا Java بر روی کامپیوترتان راه اندازی شود.

Dotlet سرور نیست اما Java applet است. هر چیزی که Dotlet انجام می دهد، در کامپیوتر شخصی شما اتفاق خواهد افتاد نه در سرور EMBnet. این کار برای شما سه پیامد خواهد داشت:

- ✓ شما می توانید بعد از یکبار دانلود کردن Dotlet، بصورت offline از آن استفاده کنید. این مسئله زمانی که شما باید یک laptop که دائم به اینترنت وصل نیست کار میکنید مفید است.
- ✓ سرعت برنامه به کامپیوتر شخصی شما بستگی دارد. هرچه کامپیوترتان سریع تر باشد، Dotlet هم سریع تر کار می کند. اهمیت این موضوع یا توالی هایی بیش از ۱۰۰۰ کاراکترروشن تر می شود.
- ✓ هنگامیکه شما از Dotlet استفاده می کنید، شما توالی هایتان را به هیچ کجا نمی فرستید! هر توالی که شما با استفاده از Dotlet مقایسه می کنید، بر روی کامپیوترتان می ماند و به سرور EMBnet فرستاده نمی شود. اگر شما در یک شرکت کار می کنید که تمام توجهش در نگه داشتن اطلاعات انحصاری است، خواهید فهمید که این مزیت خیلی خوبی است!

برای آنکه مطمئن شوید هنگامی که با Dotlet کار می کنید این برنامه با هیچ سروری ارتباط ندارد، بعد از اینکه Dotlet را دانلود کردید از منوی File گزینه Work offline را انتخاب کنید.



شکل ۱-۸: صفحه Dotlet

وارد کردن توالی ها در Dotlet

در ادامه مراحلی لیست شدند، ما دو توالی که شامل یک دومین کیناز هستند را با هم مقایسه می کنیم. پیش فرض ما برای انجام این مراحل اینست که شما Dot let را دانلود کرده اید، و این برنامه در کامپیوتر شما فعال است. (در غیر اینصورت، قبل از انجام مراحل زیر به آدرس www.isrec.isb-sib.ch/java/dotlet/Dotlet.html مراجعه کنید):

۱. یک صفحه جدید اینترنت با فشار دادن همزمان دکمه های Ctrl+N باز کنید.

شما به دو صفحه جدید احتیاج دارید، به همین دلیل ما با این مرحله شروع کردیم.

۲. آدرس www.espasy.ch/cgi-bin/get-sport-fasta?p05049 را در صفحه ای که باز کردید تایپ کنید و سپس دکمه Enter یا Return را فشار دهید.

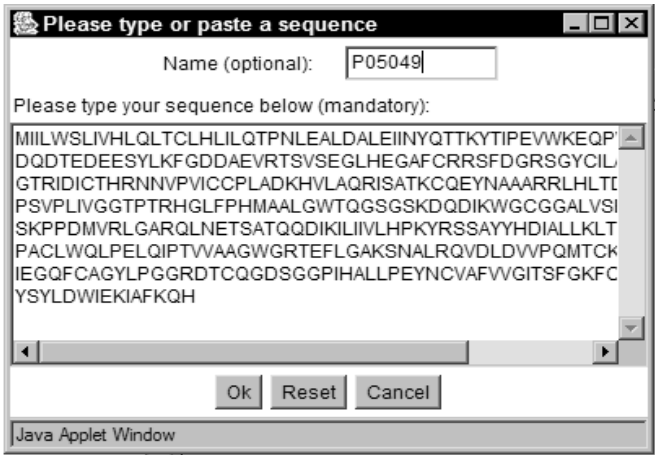
آدرس اینترنتی را به دقت تایپ کنید، توجه کنید علامت ؟ را فراموش نکنید. از طریق این آدرس و در صورتی که کد شناسایی توالی خود را بدانید، می توانید بطور مستقیم به Swiss-Prot دسترسی پیدا کنید. در مثال ما کد شناسایی P05049 است. با کمی صبر کردن یک پنجره جدید ظاهر می شود، که توالی مربوط به کد شناسایی که شما وارد کردید را نشان می دهد.

۳. توالی که روی مرورگر اینترنتتان ظاهر شده است را با استفاده از Ctrl+C کپی کنید.

۴. به صفحه ابتدایی اینترنت برگردید (پنجره Dotlet در شکل ۱-۸ نشان داده شده است) و کلید Input (در بالا سمت چپ) را فشار دهید. صفحه کوچکی شبیه به شکل ۲-۸ ظاهر می شود.

۵. با استفاده از Ctrl+V توالی را در صفحه ظاهر شده وارد کنید و سپس یک عنوان وارد کنید.

شما باید توالیتان را (فقط توالی را) به فرمت خام وارد کنید، همانطور که در شکل ۲-۸ آمده است. پنجره Name جایی است که می توانید عنوان توالی تان را در dot plot وارد کنید.



شکل ۲-۸: پنجره توالی در Dotlet

۶. کلید OK را در پنجره توالی فشار دهید.

هنگامیکه کلید OK را فشار دادید، پنجره محو می شود، این بدین معنی است که توالی تان هم اکنون در Dotlet بارگذاری شده است.

اگر شما می خواهید یک توالی را تنها با خودش مقایسه کنید، مستقیماً به مرحله ۱۳ بروید.

۷. آدرس www.expasy.ch/cgi-bin/get-sport-fasta?po5049 را در صفحه جدیدی که در مرحله ۱ باز کردید تایپ کنید و Enter یا Return را فشار دهید. این دومین توالی است که می خواهید در Dotlet بارگذاری کنید.

۸. توالی دوم مورد نظرتان را کپی کنید.

۹. در پنجره Dotlet (به شکل ۲-۸ مراجعه کنید) کلید Input را فشار دهید.

۱۰. توالی را با استفاده از ctrl+v در صفحه کوچک ظاهر شده وارد کنید، و یک عنوان برای آن بنویسید.

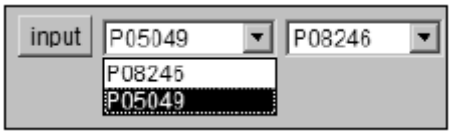
۱۱. کلید OK را در صفحه کوچک فشار دهید.

هنگامیکه کلید OK را فشار دادید، پنجره محو می شود و این بدین معنی است که توالیتان هم اکنون در Dotlet بار گذاری شده است.

۱۲. به صفحه اصلی تان برگردید (پنجره Dotlet). دو توالی که می خواهید با هم مقایسه کنید را از دو پنجره ای که نزدیک کلید Input قرار دارند انتخاب کنید (شکل ۳-۸).

شکل ۳-۸ به شما نشان خواهد داد چگونه توالی که می خواهید با هم مقایسه کنید را انتخاب کنید. اسامی که در منو هستند همانهایی هستند که شما در مراحل ۵ و ۱۰ در پنجره Title وارد کردید.

- ✓ توالی که شما در پنجره اول انتخاب می کنید، افقی قرار می گیرد (در بالای dot plot از چپ به راست)
- ✓ توالی که در منوی دوم انتخاب کردید عمودی قرار می گیرد (سمت چپ dot plot از بالا به پایین)



شکل ۳-۸: پنجره انتخاب توالی در Dotlet

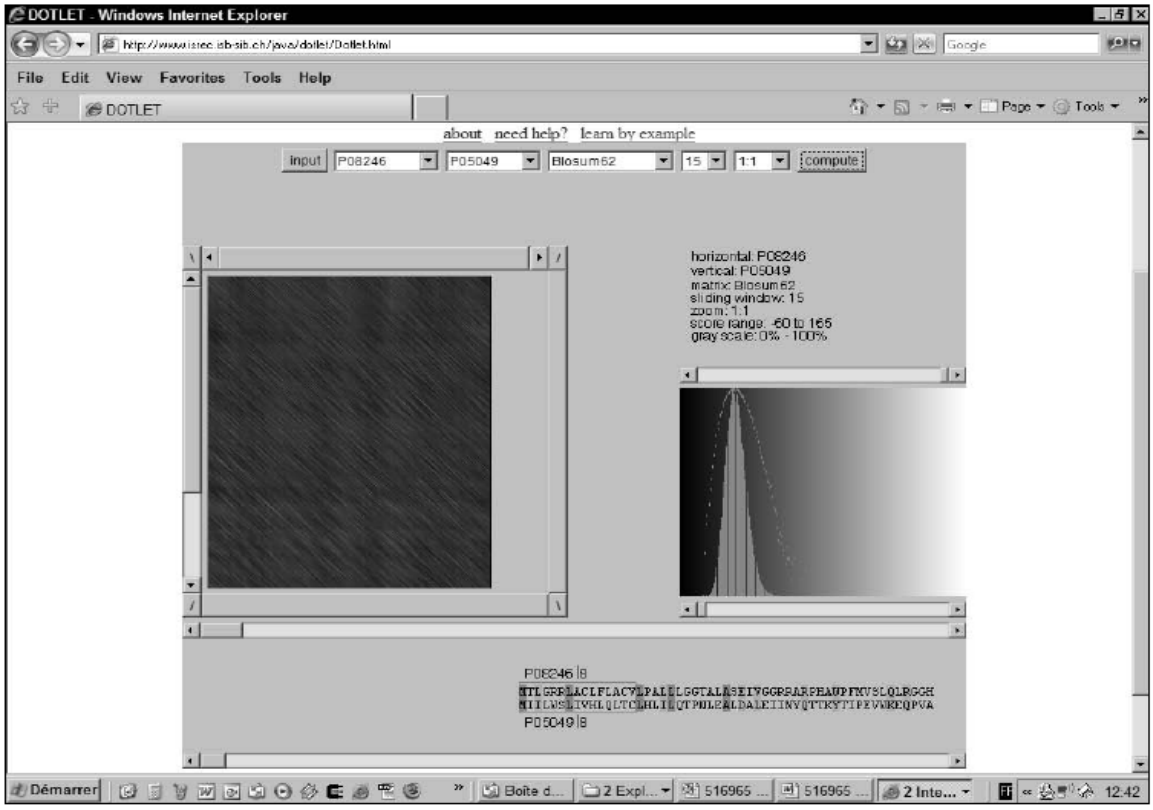
۱۳. در پنجره Dotlot کلید Compute را در انتهای سمت راست فشار دهید.

Dotlet هیچ گاه به صورت اتوماتیک کار نمی کند (بجز وقتی که شما آستانه را تغییر دهید). هر گاه شما یکی از پارامترها را تغییر می دهید ، شما باید Dotlet را با فشار کلید Compute آگاه کنید.

خروجی ، شبیه آن چیزی است که شما در شکل۴-۸ می بینید . این شکل سه پنجره Dotlet را نشان می دهد:

- ✓ پنجره dot-plot: بزرگترین پنجره در سمت چپ شامل dot-plot است.
- ✓ پنجره آستانه: این پنجره کوچک در سمت راست برای اینست که شما حساسیت dot-plot تان را تنظیم کنید.
- ✓ پنجره alignment : این پنجره نازک دراز است که ذیل آن چیزی است که نشان می دهد کدام جفت از باقی مانده ها (residues) از دو توالی، در پنجره dot-plot با یکدیگر جفت می شوند.

تفسیر کردن خروجی پیش فرض Dotlet معمولا مشکل است. برای اینکه این خروجی بیشتر مفهوم پیدا کند، باید آنرا تصفیه کنیم! ما اینها را در قسمت بعدی به شما نشان خواهیم داد.



شکل ۴–۸: فرم پیش نویس نتیجه Dotlet

Dot let تان را درست تنظیم کنید

ترتیب مراحلی که در این قسمت بیان شده است لزوماً بهترین ترتیب نیست، اما از یک رویکرد بالا – پایین پیروی می کند. ما مراحل را از مواردی که شما به سختی می توانید تنظیم کنید شروع می کنیم و با مواردی تمام می کنیم که شما باید بخوبی تنظیم کنید تا به یک dot-plot دست یابید که بیشترین اطلاعات را در اختیار شما قرار دهد.

۱. میزان بزرگنمایی را تنظیم کنید.

بطور پیش فرض، میزان بزرگنمایی Dotlet، ۱ : ۱ است. یعنی ۱ پیکسل مطابق ۱ residue است. اگر توالیتان طولانی است ، شما ممکن است همه dot plot را یکجا نبینید. این دقیقاً همان چیزی است که در شکل ۴–۸رخ داده است .

برای اینکه همه dot-plot را ببینید طبق زیر عمل کنید:

a. در منوی Zoom selector (در سمت چپ کلید Compute) ۲ : ۱ را انتخاب کنید (شکل ۵–۸).

b. کلید Compute را فشار دهید.

Dotlet هم اکنون یک dot-plot از دو توالی شما نشان می دهد.

دیدن همه ی تصویر مفید است، حتی اگر کمی کوچک باشد. بعد از اینکه شما یک دید کلی از مناطقی که در dot-plot تان جذاب هستند پیدا کردید، می توانید بزرگنمایی را به میزان اولیه برگردانید.

۲. اندازه پنجره را انتخاب کنید.

هنگامی که Dotlet دو اسیدآمینه (یا نوکلئوتید) را مقایسه می کند، اسیدآمینه (یا نوکلئوتید) مجاورش را نیز تاحدی مقایسه می کند. window size حد این مجاورت را کنترل می کند. (قسمت "ایجاد اندازه درست پنجره" را برای ببینید). برای اینکه از مثال ما استفاده کنید، باید ارزش های پیش فرض را طبق مراحل زیر تغییر دهید:

a. عدد ۵۱ را از منوی پایین رونده Window (که در سمت چپ zoom selector قرار دارد) انتخاب کنید

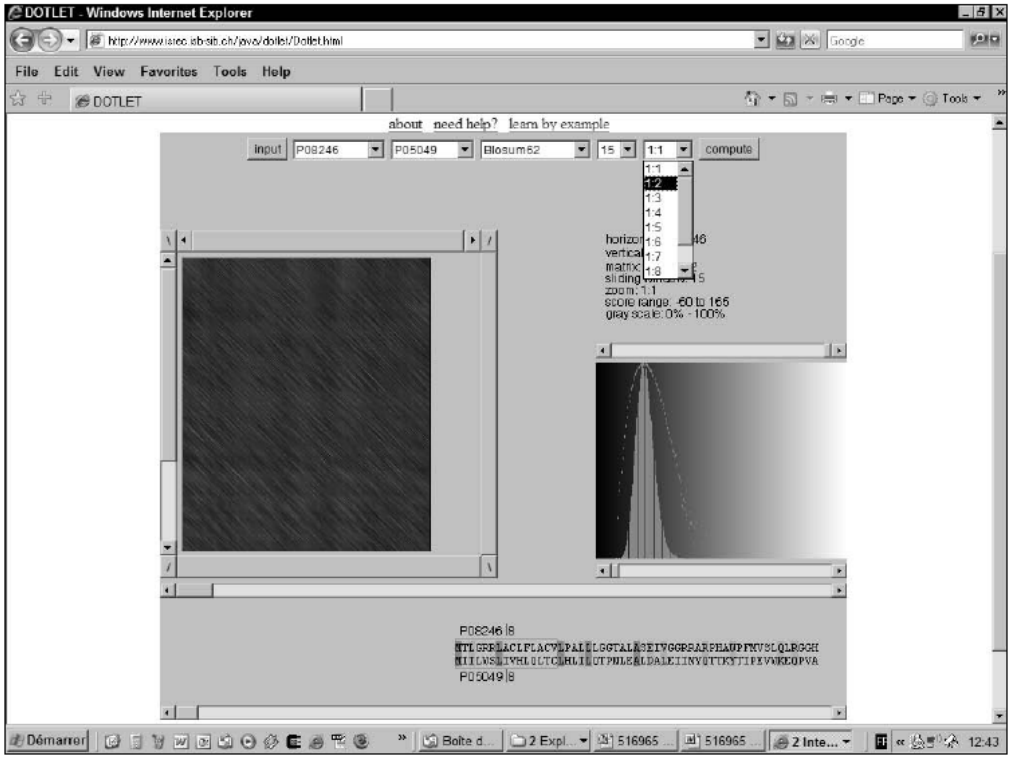
.

b. کلید Compute را فشار دهید.

بعد از تغییر دادن اندازه پنجره، شما باید آستانه را هم تنظیم کنید (مرحله ۳ را ببینید). اگر بر این اساس آستانه را تغییر ندادید، تغییر اندازه اثر چندانی نخواهد داشت.

۳. تنظیم آستانه

آستانه عددی است که رنگ هر dot را در پنجره dot-plot تعیین می کند. Dot های با نمره بالای آستانه سفید هستند. آنهاییکه نمره زیر آستانه دارند، سیاه خواهند بود. طی ساخت dot-plot، آستانه قدرتمند ترین و ظریف ترین پارامت است. خوشبختانه Dotlet به شما اجازه می دهد آستانه را بطور دینامیک تنظیم کنید و اثر آنرا همان لحظه ببینید (حتی احتیاجی به کلیک کردن روی دکمه Compute ندارید!).



شکل ۵-۸: انتخاب میزان بزرگنمایی درست در Dotlet

اندازه درست پنجره

اینجا چند قانون آسان را برای ایجاد اندازه پنجره ها ارائه می کنیم:

- ✓ پنجره های طولانی dot-plot خوبی را ارائه می دهند. آنها اطمینان ایجاد می کنند که دو اسید آمینه (یا نوکلئوتید) مشابه تنها زمانی یک dot ایجاد می کنند که اسیدهای آمینه (یا نوکلئوتید های) اطراف آن ها نیز مشابه باشند. Gurus می گوید که پنجره های طولانی dot-plot را سخت گیرانه تر ایجاد می کنند. البته dot plot نباید خیلی هم سخت گیرانه باشد. برای اینکه به تناسب درستی دست پیدا کنیم، می توان از دستور ساده زیر استفاده کرد:
- ✓ اندازه پنجره باید در همان محدوده ی اندازه عناصری باشدکه شما بدنبالش آن ها می گردید. برای مثال، اگر شما

بدنبال دامین های حفاظت شده در پروتئین خود هستید، اندازه ۵۰ اسید امینه ای یا بیشتر مناسب است.

- ✓ پنجره های کوچکتر حساس تر هستند اما نویز دارند. اینها زمانی که شما بدنبال دامنه های پروتئینی میگردید که ارتباط کمی با هم دارند، می توانند به شما کمک کنند.
- ✓ راحت تر این است که با یک پنجره بزرگتر شروع کنید، سپس آنرا کمی کوچک کنید تا سیگالی که بدنبالش می گردید ظاهر شود.

در بالای پنجره آستانه (پنجره ی سمت راست صفحه نمایش Dotlet در شکل ۴-۸) یک scroll bar کوچک افقی و یک مکان نما وجود دارد. این مکان نما، "مکان نمای بالایی" است. در پایین همین پنجره یک scroll bar کوچک افقی دیگر به همراه یک مکان نمای دیگر وجود دارد که به آن "مکان نمای پایینی" می گویند.

محور افقی در پنجره آستانه نمره را نشان می دهد. نمره های پایین در سمت چپ، نمره های بالا در سمت راست هستند. پیک بزرگ نمودار به شما معمول ترین نمره دو residue را، که به صورت تصادفی انتخاب شده است، نشان می دهد.

شما می توانید آستانه را بصورت دو خط عمودی که از مکان نمای بالایی و پایینی عبور می کنند، فرض کنید. موقعیت آستانه ظاهر پنجره dot-plot را کنترل می کند (پنجره بزرگ در چپ شکل ۴-۸).

در این پنجره dot-plot:

- نقاط با نمره زیر آستانه مکان نمای پایینی سیاه هستند.
- نقاط با نمره بالای آستانه مکان نمای بالایی سفید هستند.
- نقاط با نمره بین آستانه دو مکان نما خاکستری هستند.

همین تنظیم آستانه است که باعث ایجاد الگوی جالبی از residue ها بصورت نقاط سفید در زمینه سیاه در پنجره dot-plot می شوند. شما می توانید از روش زیر برای رسیدن به این اثر استفاده کنید:

a. مکان نمای پایینی در پنجره آستانه را کاملاً به سمت راست بکشید.

این باعث می شود dot-plot در سمت چپ کاملاً سیاه شود.

b. سمت راست پنجره آستانه را از وسط بگیرید و در حالیکه کلید راست موس را گرفته اید به آرامی آنرا به چپ بکشید.

در حینی که آنرا می کشید هنگامی که مکان نمای بالا و پایین روی یکدیگر هستند، نقطه ها در dot-plot سیاه یا سفید هستند. اگر یکی از مکان نما ها را به سمت چپ یا راست حرکت دهید، نقاطی که نمره آن ها بین آن دو آستانه باشد، به رنگ خاکستری در می آیند. به این ترتیب نمودار برای نمایش عمومی زیباتر خواهد شد.

اگر می خواهید نقاط سیاه در پس زمینه سفید ببینید، مکان نمای بالایی را به آرامی به سمت راست مکان نمای پایینی تغییر مکان دهید. به این ترتیب رنگ ها بر عکس می شوند.

۴. dot-plot را ذخیره کنید.

مهم ترین قسمت استفاده از Dotlet ، ذخیره کردن نتایج است. همه مرورگر های انترنت نمی توانند این نتایج را پرینت کنند و ما نمی توانیم مرورگری را که بتواند تمام نتایج مشاهده شده را در قالب یک فایل ذخیره کند پیدا کنیم. تا آنجا که ما می دانیم تنها دو راه وجود دارد که بتواند مشاهدات را ثبت کند:

- ✓ محتویات browser را در فایل pdf پرینت کنید. برای اینکار باید یک PDF printer در کامپیوترتان نصب کنید.

- ✓ از تکنیک ol'screen-capture استفاده کنید. شما می توانید وقتی برای ذخیره هر نوع نمایشگر گرافیکی دچار مشکل شدید، از این تکنیک استفاده کنید. از روش های زیر برای انجام screen capture (عکس گرفتن از کل صفحه) استفاده کنید.

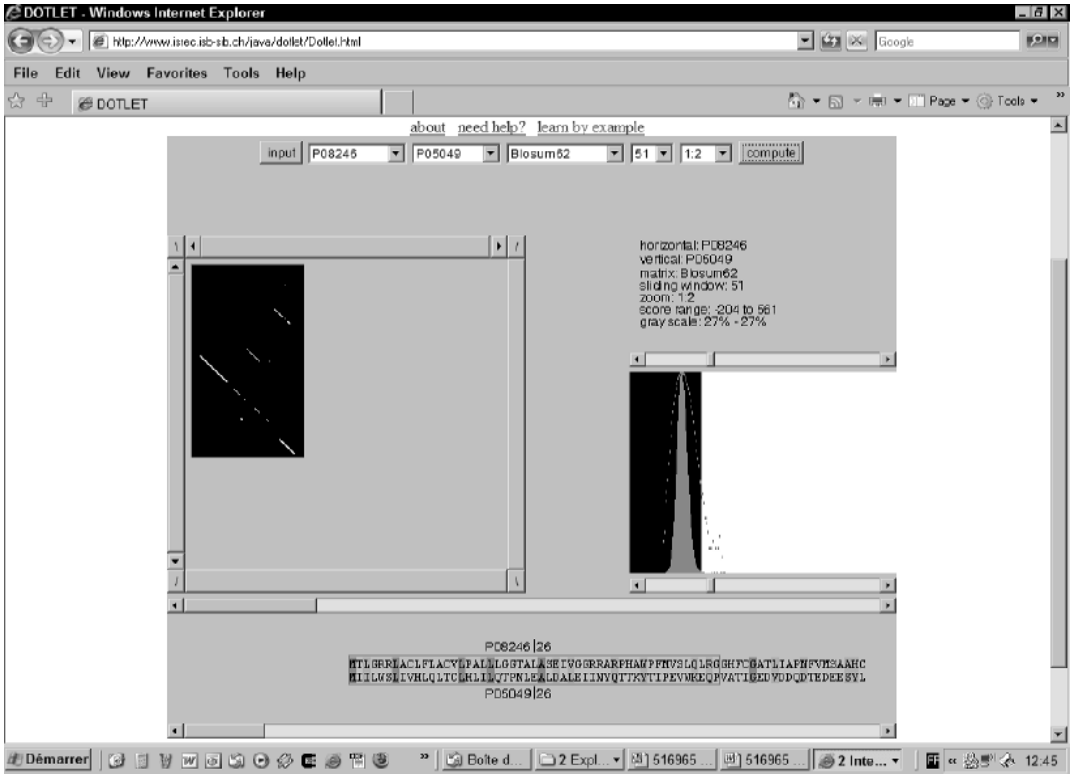
a. کلید PrntScrn (معمولاً در بالا سمت راست کیبورد) را فشار دهید.
b. یک فایل PowerPoint باز کنید.
c. با استفاده از Ctrl+V عکس حاصل را به PowerPoint منتقل کنید.

هم اکنون می توانید dot-plot را پرینت کنید یا آنرا در هر نوع پوشه ای وارد کنید.

تفسیر نتایج

P05049 و P08246 (دو پروتئینی که ما در قسمت "وارد کردن توالی ها در "Dot Let در لیست مراحل استفاده کردیم) ارتباط بسیار کمی دارند. اگر شما از BLAST برای پیدا کردن یکی از دیگری استفاده کنید، شما یک تطابق با E-value برابر 10^{-4} بدست می آورید. اگر فصل ۷ (که ظرافت های BLAST را به شما یاد می دهد) را خوانده باشید، اکنون می دانید که چنین تطابقی زیاد هم خوب نیست و حتی می تواند مثبت کاذب باشد.

P05049 یک سرین پروتئاز است، بنابر این چک کردن فعالیت پروتئازی P08246 در اینترنت می تواند جالب باشد. مشکل اینجاست که وقتی شما یک آزمایش گران را بر اساس این تطابق ضعیف BLAST شروع کنید، در واقع ریسک بزرگی انجام داده اید. اگر شکست خوردید، رئیستان سوالات مرگ آوری از شما خواهد پرسید، "حواست کجا بود!" . به همین دلیل باید قبل از انجام این آزمایش مطمئن باشید که P05049 و p08246 اصالتاً در ناحیه سرین پروتئاز همولوگ هستند.



شکل ۶-۸: dot let با تنظیم مناسب پارامترها

انجام dot-plot راه مناسبی برای پاسخ به این پرسش است. ایجاد dot-plot شبیه به آنچه ما در مراحل قبلی لیست کردیم بسیار بجا و به موقع است. شکل ۶–۸ نمونه مناسبی برای وجود یک دامین محافظت شده بین دو پروتئین است. شما می توانید از این امر با نگاهی انداختن به حاشیه نگاری Swiss-Prot مطمئن شوید (از طریق www.expasy.ch در دسترس است). Swiss-Prot می گوید که P05049 و P08246 هر دو شامل یک دامین سرین پروتئازی، در همان ناحیه ای که dot plot نشان داد، هستند. البته در حالت واقعی شما ممکن است حاشیه نگاری Swiss-Prot را برای دو پروتئین پیدا نکنید. در واقع ، در حالت واقعی ممکن است اصلاً هیچ حاشیه نگاری نداشته باشید! این نباید شما را از ایجاد dotplot باز دارد . حتی اگر عملکرد یک دامین محافظت شده را ندانید، تشخیص یک دامین بین دو توالی پروتئینی راه مناسبی است برای یافتن ناحیه خاصی که مسئول عملکرد پروتئین است.

.

انجام آنالیزهای بیولوژیک با dotplot

مثال قبلی نشان می دهد که چگونه دامینی را که در دو پروتئین ارتباط کمی با هم دارند را پیدا کنیم. Dotplot ها نمایشهای بسیار شاخصی از ارتباطات بیولوژیکی ارائته می دهند. در استفاده از آن ها کافی است تعدادی از الگوهای گرافیکی را بشناسیم و بدانیم هر کدام به چه موقعیت هایی تعلق دارند.

در مثال بعدی، ما به شما بعضی از این موقعیت ها را به شما نشان خواهیم داد. برای هر یک، ما فرض می کنیم شما توالیتان را در dot let با استفاده از پروسه ای که ما در قسمت " استفاده از DOT LET در اینترنت" برای شما شرح دادیم ، وارد کرده اید.

تشخیص توالی های پشت سر هم

پروتئینها معمولا شامل یک دامین کوچک هستند که بارها تکرار شده است. بنظر می رسد که دو برابر شدن های متوالی ابزاری هستند که معمولا برای ایجاد تغییرات تکاملی استفاده می شود تا پروتئین های جدیدی بسازد یا فعالیتشان را کارآمدتر کند . معمولاً دامین هایی که اغلب به در توالی تکرار می شوند شامل دامین فیبرونکتین[1] (که ما بیشتر دوست داریم آن را "molecular Velcro" بنامیم) ، EF-hands که به کلسیم متصل می شوند، یا انگشتهای Zn [2] که در DNA bindig نقش دارند، می شوند.

Dotplot بهترین راه برای مکان یابی دامین های تکراری هستند. در شکل ۷–۸ توالی پروتئین انسانی YE73 نشان داده شده است. YE73 یک فاکتور بالقوه نسخه برداری انسانی است (در نسخه برداری RNA شرکت می کند) ، که از ۱۳ دامین بسیار محافظت شده انگشت Zn پشت سر هم تشکیل شده است.

برای آنالیز این توالی طبق مراحل زیر عمل کنید:

۱. وارد آدرس اینترنتی www.expasy.ch/cgi-bin/get-sport-fasta?po5049 شوید.
Q9P255 شماره شناسایی برای توالی پروتئین انسانی YE73 است.
۲. توالی را در Dotlet وارد کنید (طبق مراحلی که در قسمت "استفاده از Dotlet در اینترنت" به آن اشاره کردیم).
۳. آستانه و تنظیمات پنجره را طبق شکل۷–۸ تنظیم کنید.

شما به یک پنجره با اندازه ۲۱، بزرگنمایی ۱:۲ و مقیاس خاکستری برای مکان نمای بالا و پایین پنجره آستانه را روی ۳۹٪ تنظیم شده باشد. الگوی شکل ۷–۸ یک الگوی معمول تکرارهای پشت سرهم است. به خصوصیات زیر توجه کنید:

- ✓ خط مورب اصلی dot plot توالی را در مقابل خودش را نشان می دهد.
- ✓ تکرارها بصورت خطوط مورب بلند پیوسته ای بالا و پایین خط مورب اصلی نمایش داده می شوند.

[1] Fibronectin
[2] Zn fingers

- ✓ خطوط مورب دارای فواصل منظم هستند.
- ✓ مجموع خطوط مورب یک شکل مربع را تشکیل داده اند.

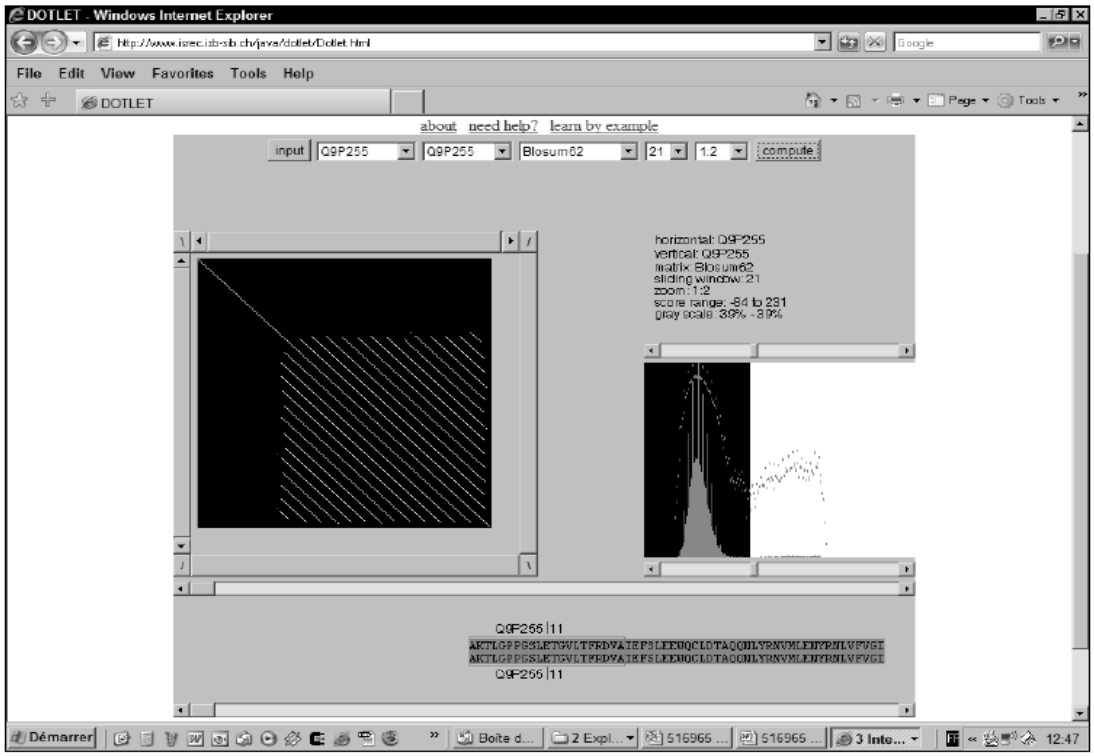
هر گاه چنین الگوی گرافیکی را دیدید، بدانید در حال نگاه کردن به قطاری از تکرارها هستید. در این صورت می توانید از این تجمع قطرها سه نتیجه جذاب بگیرید:

- ✓ تعداد تکرارها برابر است با تعداد خطوط مورب بالا (یا پایین) خط مورب اصلی به اضافه ی خط مورب اصلی.
- ✓ اگر هیچ فاصله گذاری (spacing)، بین تکرار ها نباشد، فاصله بین دو قطر مجاور اندازه تکرار را نشان می دهد.
- ✓ کوتاهترین خط مورب مختصات هماهنگی یک تکرار را به شما می دهد.

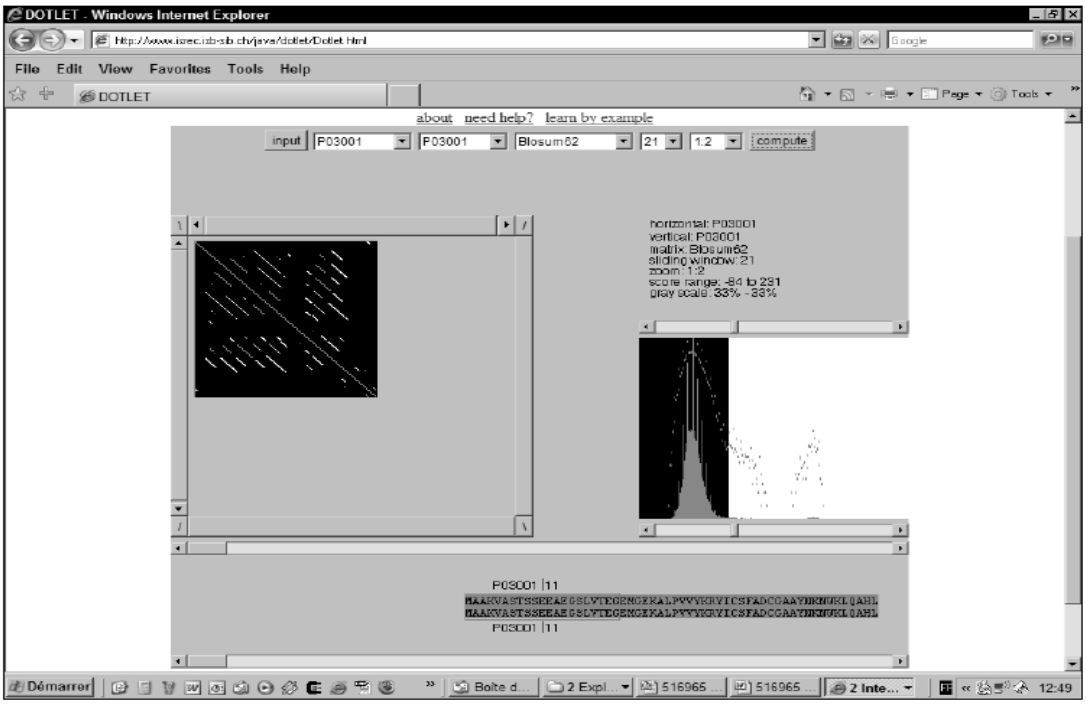
YE73 یک مثال فوق العاده است، زیرا از توالی های بسیار محافظت شده تشکیل شده است. اما زندگی همیشه به این سادگی نیست و در بسیاری از موارد، دامین های تکراری(به دلیل تجمع جهش ها) به این اندازه شبیه به هم نیستند. پروتئینهایی که شامل تکرارهای با ارتباط کم هستند، الگوهای dot-plot پیچیده تری ایجاد می کنند. در این پروتئینها، هر واحد تکرار شونده لزوماً تمام واحدهای دیگر را شناسایی نمی کند.

شکل ۸-۸ چنین موردی را به شما نشان می دهد. Tf3a یک فاکتور رونویسی است که هم چنین از دامین تکرار شونده انگشت Zn تشکیل شده است. شما می توانید توالی را در آدرس www.expasy.ch/cgi-bin/get-sport-fasta?PO 3001 بیابید.

Tf3a داستان دامین های انگشت Zn را که بیان می کند که یک زمانی با پدیده دو برابر شدن از هم جدا شده بودند. این انگشت های Zn کمتر محافظت شده هستند و الگویی با نظم کمتر از آنچه در YE73 دیدیم نشان می دهند.



شکل ۸-۷: دامنه های تکرار شونده در Dotlet



شکل ۸-۸: دامنه های تکرار شوند با محافظت کم Tf3a در Dotlet

دامنه های پشت سر هم تنها فرم دامنه های تکراری در پروتئین یا DNA نیستند. می توانید با خیال راحت مطمئن باشید که هر نوع از تکرار که شما می توانید به آن فکر کنید احتمال دارد در پروتئین یا DNA اتفاق بیفتد. حتی اگر این موضوع تا کنون کشف نشده باشد، می توانید بر سر این شرط بندی کنید که قرار است کشف شود!

اثر یک تکرار معمولا یکسان است: این اثر یک خط مورب است که هنگامیکه شما توالی را با خودش مقایسه می کنید دور تر از خط مورب اصلی قرار می گیرد.

دامنه ی تکرار شده می تواند به شما کمک کند تا عملکرد پروتئین را روشن سازید. اگر دامنه تکرار شده ای بدون عملکرد مشخص و بدون هیچ شباهتی با هیچ کدام از پروتئین های خاص، پیدا کردید، می توانید طبق موارد زیر عمل کنید:

۱. هر واحد تکراری را جدا کنید.
۲. یک multiple sequence alignment از دامین ها ایجاد کنید (فصل ۹ را ببینید).
۳. موقعیت های محافظت شده را در دامین پیدا کنید.
۴. فرم دامین را به الگو یا پروفایل PROSITE تغییر دهید.
فصل ۶ را برای بیشتر دانستن راجع به الگو های PROSITE ببینید.
۵. Swiss-Prot را جستجو کنید تا ببینید آیا این الگو با عملکرد خاصی مرتبط است یا خیر؟

پیدا کردن نواحی با پیچیدگی کمتر (Low-complexity)

نواحی با پیچیدگی کم قطعاتی هستند که یک آمینو اسید خاص را بیشتر از آنچه که در یک پروتئین نرمال انتظار میرود داشته باشند. برای مثال، یک قطعه با ۱۰۰ اسید آمینه که از ۴۵ پرولین تشکیل شده باشد، مشخصاً یک قطعه با پیچیدگی کم است.

نواحی با پیچیدگی کم معمولا عملکردهای بیولوژیک دارند مثل بر هم کنش های پروتئین- پروتئین (Leucine Zipper) یا باند شدن غیر اختصاصی DNA/RNA (دامین های غنی از ARG).

خوشبختانه دیدن نواحی با پیچیدگی کم در dot-plot بسیار ساده است. هنگامیکه یک توالی را با خودش مقایسه می کنید، نواحی با پیچیدگی کم بصورت مربع هایی نمایان می شوند. می توانید یکی از اینها را در شکل ۹–۸ ببینید.
برای ایجاد نتایجی شبیه آنچه در شکل۹–۸ دیدید طبق موارد زیر عمل کنید:

۱. وارد آدرس اینترنتی www.expasy.ch/cgi-bin/get-sprot-fasta?P21997 شوید.
P21997 کد شناسایی sulfated surface glycoprotein 185 است.
۲. توالی را وارد Dotlet کنید (طبق مراحلی که در قسمت ""Using dotlet over the inter net ذکر شد).
۳. تنظیمات را همانطور که در شکل۹–۸ می بینید انجام دهید.
تنظیمات پنجره را روی ۷، میزان بزرگنمایی را روی ۱:۲ و مقیاس خاکستری برای مکان نمای بالا و پایین پنجره آستانه را روی ۴۴٪ تنظیم کنید.

آنالیز اسیدهای نوکلئیک با dotlet

در حالت طبیعی dotplot برای تهیه نقشه از ژن ها طراحی شده است. اما متأسفانه، dotplot کمی برای چنین عملکرد هایی که به توالی های با بیش از ۱۰۰۰۰ کاراکتر سروکار دارند، کند است.
Dotlet برای آنالیز توالی های اسید نوکلئیک شامل دو ویژگی مفید است:

- ✓ اگر یک توالی اسید نوکلئیک و توالی پروتئینی را مقایسه می کنید : Dotlet به صورت اتوماتیک توالی نوکلوئید را به سه قالب باز خواندن ممکن ترجمه می کند. با این روش شما می توانید مرزهای اینترون–اگزون را به دقت مشخص کنید .
- ✓ اگر توالی نوکلوئیدی را با خودش مقایسه می کنید : Dotlet به صورت اتوماتیک جای یکی از توالی ها را با توالی مکملش عوض می کند. بدین ترتیب، اگر توالی شما حاوی دو رشته مکمل باشد (مثل ساقه سنجاق سر در ساختار یک RNA)، اینها بصورت خطوط موربی ظاهر می شوند که از گوشه راست بالا به سمت چپ پایین می روند (عمود بر قطر اصلی).

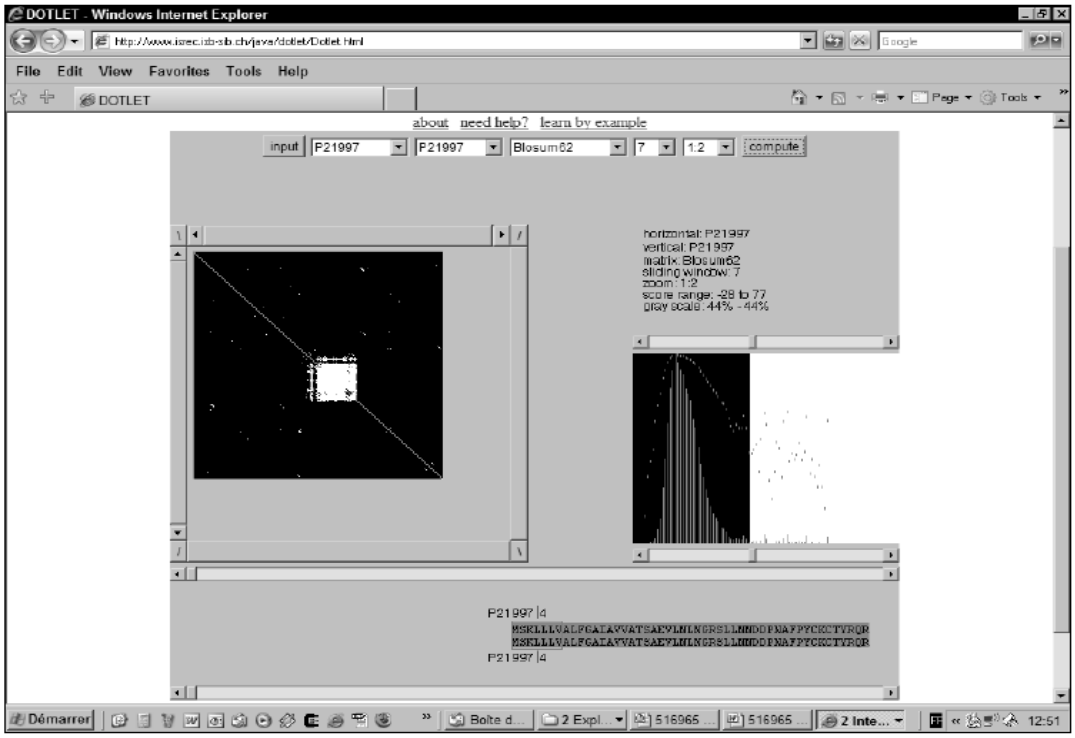
اگر شما می خواهید از Dotlet برای آنالیز های خالص تر توالی ها استفاده کنید، ما به شما یک جزوه آموزشی بسیار خوب آنلاین که توسط مؤلفین Dotlet تولید شده (www.isrec.isb-sib.ch/java/Dotlet-examples.html) را پیشنهاد می کنیم.

ساخت Local Alignment در اینترنت

اگر سری به قسمت dot-plot در این فصل بزنید، حتما به این موضوع پی خواهید برد که ایجاد dot-plot یک راه ایده آل برای فهم بهتر ارتباط بین توالی ها است . با این وجود ، dot-plot قابل پیش بینی نیست ، چراکه با وجود نشان دادن سیگنال، نمی توان تفسیر مشخصی از آن داشت. بنابراین به روشی برای ایجاد یک alignment (هم ترازی) مناسب احتیاج دارید.

دو نوع هم ترازی (یا به اصطلاح alignment) وجود دارد : global (هنگامیکه دو توالی در تمام طول خود هم تراز می گردند) و local (هنگامیکه برنامه تنها قسمت های مشابه دو توالی را هم تراز می نماید و مابقی را نادیده می گیرد).

اگر برنامه dot-plot ، نشان دهنده ارتباط دو توالی در تمام طول می باشد، دیگر احتیاجی به ساخت local alignment نیست و می توانید مستقیماً به سراغ قسمت global alignment دراین فصل بروید. اما اگر اطمینان ندارید، لطفا همین جا بمانید!



شکل ۹-۸: شناسایی قطعات با پیچیدگی کم با Dotlet

دو دلیل اصلی برای آنالیز توالی های با هم ترازی های Local:

- ✓ به منظور مقایسه دو توالی کاملاً مرتبط با تنها تعداد اندکی domain غیر مجاور
- ✓ برای آنالیز موارد تکرار شونده موجود در یک توالی

روش local alignment ، همانطور که از نام آن مشخص است، یک هم ترازی از نواحی مشابه دوتوالی به شما ارائه می دهد.

نکته حائز اهمیت در مورد این نوع از هم ترازی ها، حذف اتوماتیک اسید های آمینه و نوکلئوتید های غیر هم تراز می باشد.

در تئوری، یک local alignment باید با یکی از قطرهای ظاهر شده در dot-plot تطابق داشته باشد . با این حال در عمل کار ساده ای نیست، چراکه ممکن است سیگنالی در dot-plot مشاهده نمایید که با هیچ local alignment مطابقت ندارد و گاهی برنامه های local alignment ، هم ترازی هایی را گزارش می کنند که در dot-plot ظاهر نمی شود. بنابر این تنها در صورت هم خوانی dot-plot و هم ترازی می توان از نتیجه کار اطمینان داشت.

قانون دیگر در مورد هم ترازی توالی ها اینست که شما تنها باید در صورتی به نتایج حاصله اطمینان داشته باشید که واقعا و به وضوح از صحت آنها آگاهی یابید. بنابراین تنها در صورت وجود مستندات و اطلاعات ساختاری و آزمایشگاهی مطمئن می توانید از یک هم ترازی مشکوک استفاده کنید

انتخاب **global alignment** مناسب

دو روش برای ایجاد هم ترازی های global وجود دارد : روش سریع ذهنی به نام BLAST و یک روش کندتر ، اما کارآمدتر به نام Lalign . نقاط ضعف و قوت این دو روش در جدول ۳-۸ لیست شده است، آن را به طور دقیق مطالعه نمایید و از بین این دو، روشی مناسب تر را برای انجام کار خود انتخاب کنید.

ورژنی از برنامه BLAST که به منظور مقایسه دو توالی استفاده می گردد، در حقیقت همان ورژنی است که می توان برای جستجوی یک پایگاه اطلاع رسانی نیزاز آن استفاده کرد. می توانید برنامه را به دو توالی محدود نمایید، هم ترازی حاصله با نتایج حاصل از کار با BLAST در صورت جستجوی کامل پایگاه های اطلاع رسانی ، تفاوتی نخواهد داشت . بطور خلاصه ، استفاده از BLAST مطلبی بیش از آنچه ازجستجوی یک پایگاه اطلاع رسانی به آن پی خواهید برد به شما ارائه نمی دهد. می توانید از طریق آدرس زیر به این نوع خاص از BLAST دست یابید:

www.ncbi.nlm.gov/blast/bl2seq/wblast.cgi

Property	*BLAST*	*Lalign*
Speed	Very fast	Slower
Size of sequences you can use	Very long sequences	Shorter sequences
Score	E-value	Matrix score and E-value
Alignments	Reports the single best	Reports the ten best (or more)
Sequence type	Best with DNA	Best with proteins

جدول ۳-۸: استفاده از BLAST یا Lalign برای مقایسه جفتی

استفاده از **Lalign** برای پیدا کردن ده **global alignment** برتر

اصلی ترین مشکل پیشرو در حین استفاده از BLAST اینست که تنها یک هم ترازی و البته بهترین آنرا به ما نشان می دهد. این برای جستجودر پایگاه اطلاع رسانی بسیار مفید خواهد بود ، اما ممکن است به اندازه کافی برایتان جالب نباشد، آنهم درصورتیکه ناحیه ای از توالی که مدنظرتان بود ، یک local alignment خوب باشد و جزء بهترین local alignment نباشد . شبیه فروشگاه موزیک که تنها یک سی دی به شما می فروشد آنهم پر فروش ترین سی دی در هفته در صورتیکه این سی دی ، سی دی مورد علاقه شما نیست.

Lalign ، بهترین مکمل BLAST است. کندتر است اما با کارایی بیشتر، به طوریکه تمامی هم ترازی های Local (از بهترین آنها تا آنهاییکه در درجات پائین تری قرار دارند) را شامل می شود . نگران مسئله سرعت هم نباشید . در توالی های با کمتر از ۱۰۰۰ سمبل ، احتمالاً تفاوتی را بین سرور bl2seq و Lalign احساس نخواهید کرد .

بطور کلی ، Lalign به هنگام آنالیز پروتئین های پیچیده و پراز تکرار فوق العاده است. با انجام مراحل زیر نحوه کار با Lalign برای تشخیص local alignment از دو توالی مرتبط با یکدیگر که حاوی serine protease domain می باشند را فرا خواهید گرفت:

۱. وارد آدرس اینترنتی www.ch.embne.org/software/LALIGN_form.hyml شوید.
صفحه Lalign از ch.embne.org. ظاهر خواهد شد .
۲. Local را از گزینه های موجود در قسمت option انتخاب نمایید، همانطور که در شکل ۱۰-۸ آمده است .
می توانید از Lalign interface برای انجام local alignment و global alignment استفاده کنید .
۳. شماره ی تحت هم ترازی های گزارش شده را انتخاب کنید .

شماره ای که تقریباً مرتبط با شماره ی قطرهایی که در dot-plot دیده اید را انتخاب کنید .
اگر به Lalign برای گزارش ۱۰ local alignment احتیاج دارید، گزینه ۱۰ را انتخاب کنید .

۴. substitution matrix در حالت پیش فرض انتخاب کنید .

substitution matrix، موتاسیون ها را هنگامی که Lalign توالی ها را هم تراز می کند ، کنترل می کند. ماتریکس در حالت پیش فرض، یک ماتریکس برای DNA و یک BLOSUM45 برای پروتئین ها است ، در بیشتر موارد ، این پیش فرض کاملا مناسب است و اگر می خواهید ماتریکس را عوض کنید ، در ذهن داشته باشید که شاخص بالای BLOSUM، Lalign را غیر قابل انعطاف تر و البته دقیق تر می کند و در نتیجه هم ترازی های ایجاد شده، کوتاهتر می شود . شاخص پایین PAM اثر معکوس دارد .

۵. تنظیم gap-opening penalty

gap-opening penalty ، به منظور تعیین ارزش باز شدن یک gap در یکی از توالی ها است. اگر آنرا بالاتر از میزان پیش فرض تنظیم کنید ، local alignment هایی که شامل gap می باشد ممکن است به چندین هم ترازی کوتاهتر شکسته شوند. قانون خاصی برا ی پیش بینی optimal value برای gap penalty وجود ندارد. value های پیش فرض بر روی ماتریکس های پیش فرض راه اندازی شده اند . اگر تنظیمات ماتریکس را تغییر دهید ، Lalign نمی تواند بطور اتوماتیک value هارا دوباره تنظیم نماید . برای مثال اگر شما از یک BLOSUM با شاخص پایین تر از ۴۵ استفاده می کنید ، باید gap-opening value را کاهش دهید و آنرا کمتر از ۱۴ تنظیم کنید .

۶. تنظیم gap-extension penalty

gap-extension penalty ، یک خطای اضافه در طول gap است این خطا به خطای gap-opening اضافه شده تا نتایج مربوط به یک gap بدست آید . این مورد باید حدوداً ۱۰مرتبه کمتر از gap-extension penalty اتفاق بیافتد . هنگامیکه شما توالی های بسیار مرتبط را مقایسه می کنید ، خطای gap-opening بالا و خطای gap-extension خیلی پایین معمولاً بهترین نتیجه راارائه می دهند . توجه داشته باشید که تأکید بر تشخیص gap ها بیشتر بعلت وجودشان می باشد تا بعلت طولشان .

۷. ID یا Swiss-Port AC را از منوی **Input Sequence Format** انتخاب نمایید. برای یافتن منو به پائین صفحه Lalign مراجعه کنید.

۸. P05049 را در چهارچوب مربوط به توالی اول وارد کنید.

P05049 شماره دسترسی به مارپیچ سرین پروتئاز است. اگر شما می خواهید از توالی استفاده کنید که هیچ شماره دسترسی ندارد ، آنرا در پنجره PASTE کنید و مطمئن شوید که فرمت درستی را در مرحله ۷ انتخاب کرده اید .

۹. دومین انتخاب کننده فرمت را برای ID یا Swiss-Prot AC تنظیم نمائید .

۱۰. P08246 را درچهارچوب توالی دوم وارد کنید. P08246 شماره دسترسی به الاستاز لکوسیت انسانی (human leucocyte elastase) است .

۱۱. بر روی Run Lalign کلیک نمائید . محاسبه تقریباً باید سریع انجام شود (کمتر از ۱ دقیقه). اگر برنامه به نتیجه ای نرسید ، چک نمائید تا مطمئن شوید پارامتر درست را وارد کرده اید . Lalign نتایج را به فرمت HTML document تبدیل می کند.

۱۲. با استفاده از گزینه Save As در منوی File نتایج را در یک فایل ذخیره کنید.

تفسیر نتایج حاصل از **Lalign**

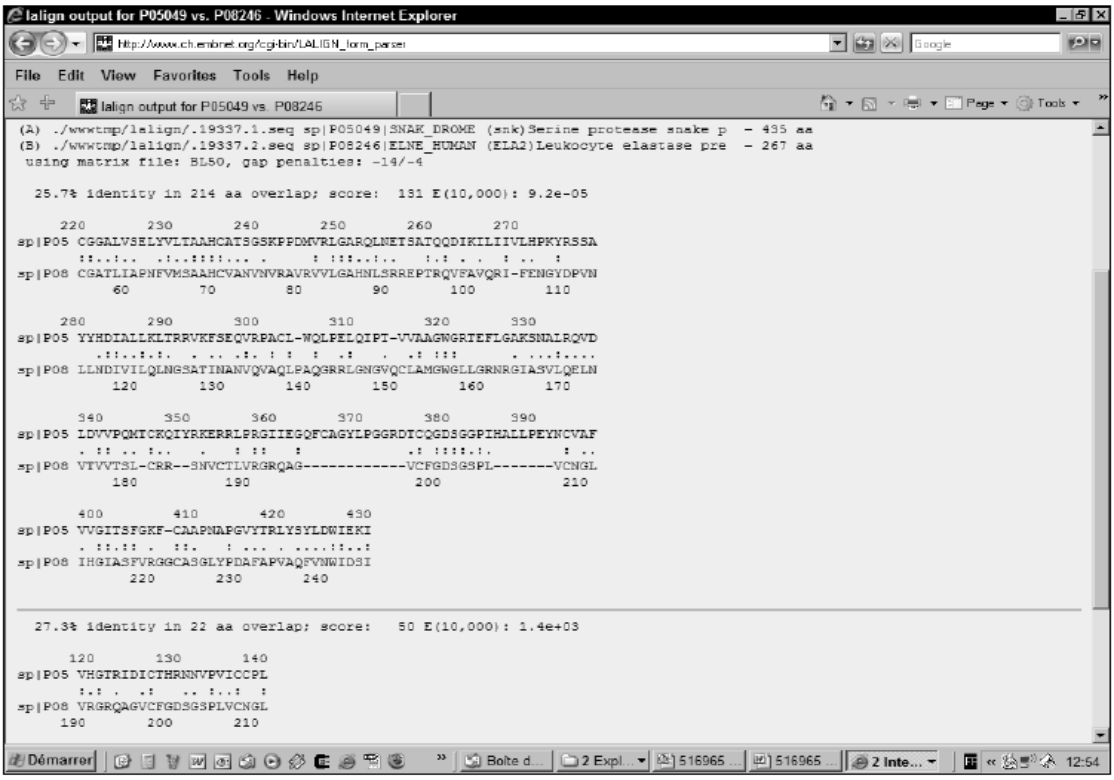
Lalign تنها تعداد هم ترازی های از پیش انتخاب شده را گزارش می کند و براساس درجه بندیشان آنها را مرتب می نماید. همانطور که در مرحله ۳ ، پیش از این نیز گفته شد یک مشخصه ی جالب Lalign اینست که تنها هم ترازی هایی را گزارش می کند که overlap نداشته باشند ، این بدان معنی است که در خروجی Lalign ، شما می توانید دو اسیدآمینه یا نوکلئوتید را تنها با یکبار هم ترازی پیدا نمائید . این دو می توانند در چندین هم ترازی ظاهر می شوند ، اما در حقیقت تنها یکبار در مقابل هم قرار می گیرند. این امر Lalign را از گزارش تفاوت های پیش پا افتاده ی یک هم ترازی Local بسیار عالی باز می دارد. شکل۱۱-۸ به شما خروجی Lalign را در مورد دو توالی که از قبل وارد کرده ایم نشان می دهد ، Lalign ردیف ها را با تنظیم کردن براساس E-Value شان در فرمت شبیه به BLAST گزارش می کند . در اولین خط بهمراه هر هم ترازی شاخصه هایی نیز بیان شده است:

- ✓ درصد همسانی: نسبت اسید آمینه های همسان که با یکدیگر هم تراز می شوند را نشان می دهد. برای مثال شما می توانید در شکل ۱۱-۸ ببینید که بهترین Local alignment ، درصد همسانی ۲۵/۷ را به خود اختصاص می دهد، دومین نمره ۲۷/۳ است .
- ✓ طول **local alignment** (Overlap) : طول کل **local alignment** را نشان می دهد .
- ✓ در جه بندی (The Score): این درجه بندی میزان gap و جانشین ها را به طور خلاصه بیان می کند، ، دقیقا همانطور که در substitution matrix و gap penalties آمده . به طور کلی ، نمره بالاتر ، هم ترازی بهتر را نشان می دهد اما توجه داشته باشید که این امر همواره صادق نیست، می توان گفت E-value شاخص بهتری برای تعیین و مقایسه کیفیت هم ترازی های (Alignment) موجود است.
- ✓ E-value: این میزان در حقیقت تعداد دفعات ایجاد چنین هم ترازی را برای دو توالی مورد نظر(البته براساس شانس) نشان می دهد. البته توجه داشته باشید که E-value بدست آمده E-value نسبت به حاصل از BLAST به هنگام جستجو در یک پایگاه اطلاع رسانی، معنا و مفهوم کمتری دارد. یک E-value خوب باید زیر 10^{-4} باشد.

Alignment، خود شامل سه نوع اطلاعات است:

- ✓ اندکس اسید آمینه در خط بالای توالی : اسید آمینه مرتبط با یک اندکس، همانی است که زیر آخرین شماره آن اندکس قرار دارد .
- ✓ یک alignment با gap هایی که توسط خطوط تیره نشان داده شده اند .
- ✓ همسانی (identity) و تشابه (similarity). در خط بین دو توالی هم تراز ، نشانه (_) نشان همسانی است ، در حالیکه (۰) نشانه تشابه است . دو اسید آمینه وقتی مشابه اند که نمره substitution هردوشان بالاتر از (صفر) ۰ باشد .

اولین هم ترازی(Alignment) گزارش شده مرتبط است با conserved serine protease domain. به خودی خود ، این هم ترازی خیلی هم متقاعد کننده نیست : شامل کمتر از ۲۶ درصد همسانی در حدوداً بیش از ۲۰۰ اسید آمینه است . برای آنکه کاملاً متقاعد کننده باشد ، ما به یک تشابه بیشتر احتیاج خواهیم داشت (تقریباً نزدیک به ۳۰ درصد) و یک E-value بهتر (کمتر) . این نتیجه ممکن است در مورد این هم ترازی درست باشد، به این دلیل که با نتایجی که ما در dot-plot دیدیم مطابقت دارد (به شکل۷-۸ مراجعه کنید). در حقیقت این امر که دو نوع متفاوت از آنالیز(dotlet & Lalign)، نتایج سازگاری به ما می دهند، ما را به وجود یک conserved serine protease domain در پروتئین می رساند.



شکل ۱۱-۸: نتیجه جستجو در Lalign

در صورت کار با پارامترهای Lalign ، حتما به این موضوع پی خواهید برد که خطا های gap-opening و یا gap-extension قویتر، می توانند قطرهای dotlet را به دو alignment (هم ترازی) بشکنند. از سوی دیگر عکس این واقعه می تواند باعث چسبیدن قطرها شود . آنچه که اتفاق می افتد در حقیقت به طول gap هایی که Lalign برای قرار دادن آنها در local alignment ها نیاز دارد وابسته است.

هنگامیکه می خواهیم در مورد کیفیت هم ترازی های(alignments) موجود قضاوت کنیم ، اندازه نقش مهمی را ایفا می نماید. اینکه قطعات کوچکی با نمرات بسیار عالی ببینید ، غیرعادی نیست . این امر می تواند از دلایل متعددی منشأ بگیرد . مثل قطعات با پیچیدگی کم (low-complexity) و یا نظم با درجه پایین در توالی های پروتئینی و از این قبیل . برای مثال ، در شکل ۱۱-۸ local alignments سوم و چهارم به وضوح بی معنی می باشند چون خیلی کوتاه هستند . هر چند ، E-value آنها بهتر از هم ترازی است که در جایگاه بالا قرار گرفته است. این موضوع به وضوح روشن می سازد که یک هم ترازی (alignment) کوتاه می تواند براحتی نمره خوبی را از روی شانس بگیرد.

متاسفانه، برنامه های بیوانفورماتیک کامل و ایده آل نیستند : ممکن است اشتباهاتی بکنند که اغلب منجر به نتایج اشتباه می شود. درست یا غلط ، این برنامه ها از یک مفهوم مطلق پیروی نمی کنند.

بهترین هم ترازی که Lalign در اینجا گزارش کرده است، محدودیت هایش را به خوبی نمایان می کند . به هر حال، به دلیل اطلاعات dot-plot ، ما اطمینان می دهیم که این هم ترازی در کل صحیح است . از سوی دیگر ، دو توالی موجود بسیار کم به هم مرتبط اند و تشابه خیلی کمی دارند . پس ما باید در تفسیرمان محتاط باشیم : بسیاری از اسید آمینه ها می توانند بطور ناصحیح هم تراز شده باشند ، به این معنی است که احتمال دارد این هم ترازی صحیح و غلط باشد . تنها راه برای اطمینان از صحت این آنالیزها اینست که یک multiple sequence alignment بسازیم یا از اطلاعات ساختاری موجود استفاده

نمائیم (برای اطلاعات بیشتر در مورد هم ترازی های توالی های چندگانه ، فصل ۹ را مطالعه نمائید و برای فهم بهتر اطلاعات ساختاری و ساختارها، به سراغ فصل ۱۱ رابروید).

ساخت **global alignment** در اینترنت

یک global alignment تنها نامی است برای بیان یک هم ترازی از تمام اسیدآمینه ها یا نوکلئوتیدهایی که در توالیتان آمده است . هنگامیکه تصمیم می گیرید تا global alignment بسازید ، شما بطور ضمنی این فرضیه را در ذهن پرورانده اید که دو توالیتان در تمام طولشان با هم مرتبطند .

global alignment ها به منظور ساخت multiple sequence alignment بسیار حائز اهمیت می باشد . با این وجود هنگامیکه تنها از دو توالی استفاده می کنید ، جذابیت کمی برای انجام آنالیزهای بیولوژیک دارند . در حقیقت به هنگام در دست داشتن دو توالی نزدیک به هم، می توانید نتایج قابل قبولی را با جمع آوری توالی های مرتبط تر بدست آورید و در نهایت یک multiple sequence alignment تولید نمائید (فصل ۹).

global alignment کاملاً برای کشف تشابهات بین دو توالی مفید نیستند، به این دلیل که روشهای آماری برای ارزیابی E-value آنها کاربردی نیست . اگر چه مبتدیان معمولا global alignment را به local alignment ترجیح می دهند(البته به دلیل فهم آسان آنها).

در یک global alignment ، هیچ اسیدآمینه یا نوکلئوتیدی بطور راز گونه محو نخواهد شد . شما در خروجی حاصل به وضوح آنچه را که در چهارچوب مربوطه قرار داده اید، خواهید دید ، البته به همراه تعداد اندکی gap.

دلایل ساخت global alignment:

- ✓ چک کردن تفاوت های کوچک بین دو توالی. این ممکن است با داده هایی که شما دستکاری کرده اید و احتمالاً آنها را تغییر داده اید، ایجاد شود . global alignment ، بهترین روش برای موقعیت یابی اشکالات احتمالی است.
- ✓ آنالیز کردن پلی مرفیسم (برای مثال ، SNPs) بین توالی های مرتبط به هم .
- ✓ مقایسه دو توالی که در قسمتی با هم overlap دارند . در این مورد ، احتیاج به انجام مقایسه global pairwise که بی نظمی انتهای توالی را در بر نگیرد خواهید داشت.

استفاده از **Lalign** برای ایجاد **global alignment**

در بحث الگوریتم ، تفاوت زیادی بین یک global alignment و local alignment وجود ندارد . در صورت تمایل می توانید برنامه ای برای هر دوشان بنویسید ، اما بیش از یک خط تفاوت بین این دو الگوریتم نخواهید یافت !

شما می توانید از Lalign برای ایجاد global alignment استفاده نمائید، اما انتخابهای دیگری هم وجود دارد که ما بعضی از آنها را در جدول ۴-۸ (در انتهای این فصل) لیست کرده ایم .

استفاده از Lalign برای ایجاد global alignment کاملاً شبیه به استفاده از Lalign برای ایجاد local alignment است . تنها مطلب اضافه ای که به آن احتیاج دارید، کلیک بر روی گزینه radio (global whithout end-gap) penalty نزدیک به بالای فرم Lalign در آدرس زیر است:

www.ch.embnet.org/software/LALIGN_form.html

هم ترازی پروتئین ها و DNA

ابزارهای هم ترازی که ما در بالا برای شما توصیف کردیم تنها زمانی بکار می آیند که می خواهید توالی هایی از یک نوع مشابه را مقایسه کند . پروتئین با پروتئین و یا DNA با DNA . گاهی ، شما احتیاج دارید تا پروتئین را با قطعه ای از DNA مقایسه کنید (برای مثال ، ژن اصلی آن) . انجام این چنین هم ترازی هایی به ابزارهای خاصی احتیاج دارد که می تواند gap های طولانی مرتبط با اینترون ها در نظر بگیرد . حداقل دو محل برای انجام این کار به صورت آنلاین وجود دارد ، یکی در انستیتو پاستور (bioweb.pasteur.fr/seqanal/interfaces/prota12dna.html) و دیگری توسط دکتر پیربورک و گروهش که در آزمایشگاه بیولوژی ملکولی اروپا در هیدلبرگ راه اندازی شده (coot.embl.de/pa12nal). هردو سرویس خدماتی از شما خواهد خواست تا پروتئین را در اختیارشان قرار دهید و DNA مرتبط با آنرا قبل از آنکه شروع به استفاده از سرویس ارائه شده نمائید، دراختیار داشته باشید.

اگر شما می خواهید تنها پروتئین را دراختیار داشته باشید ، می توانید حتی یک blastx را در مقابل genume راه اندازی کنید (فصل ۶ را ببینید)، یا اینکه از سرور تحت وب با نام protogene استفاده نمائید که بطور اوتوماتیک توالی DNA مرتبط با پروتئین داده شده را دراختیارتان قرار می دهد.

مقایسه دو به دو توالی به طور رایگان در اینترنت

تعداد زیادی سرور به منظور مقایسه دو جفتی طراحی شده اند . جدول ۴-۸ خلاصه ای از سرور های معتبر موجود را ارائه می دهد.

Name	*Address*	*Alignment Type*
Lalign	www.ch.embnet.org/software/LALIGN_form.html	Global/Local
Lalign	fasta.bioch.virginia.edu/fasta_www/plalign.htm	Global/Local
USC	www-hto.usc.edu/software/seqaln/seqaln-query.html	Global/Local/
Alion	fold.stanford.edu/alion/	Global/Local
Align	genome.cs.mtu.edu/align.html	Global/Local
Align	www.ebi.ac.uk/emboss/align/	Global/Local
xenAliTwo	www.soe.ucsc.edu/~kent/xenoAli/xenAliTwo.html	Local for DNA
Blast2seqs	www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi	Local BLAST
Protal2dna	bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html	Protein against DNA
Pal2nal	coot.embl.de/pal2nal	Protein against DNA

جدول ۴-۸: برنامه های آنلاین alignment جفتی

ایجاد یک alignment همیشه کافی نیست. گاهی ممکن است نیاز باشد این alignment بررسی و اهمیت آماری آن سنجیده شود. جدول ۵-۸ چند سایت را برای این کار به شما معرفی می کند.

Name	*Address*	*Function*
laInview	`www.expasy.ch/tools/sim-prot.html`	Visualization
prss	`www.ch.embnet.org/software/PRSS_form.html`	Evaluation
prss	`Fasta.bioch.virginia.edu/fasta/prss.htm`	Evaluation
graph-align	`darwin.nmsu.edu/cgi-bin/graph_align.cgi`	Evaluation

جدول ۵-۸: آنالیز دو به دو alignment به صورت آنلاین

فصل نهم

چگونه یک **Multiple sequence alignment** بسازیم؟

در این فصل می خوانیم:

- ✓ چگونه توالی های مورد نیاز برای ساخت یک Multiple sequence alignment را جمع آوری کنیم؟
- ✓ چگونه با استفاده از برنامه های ClustalWو MUSCLE یک Multiple sequence alignment بسازیم؟
- ✓ چگونه با استفاده از برنامه ی t-coffee ، Multiple sequence alignmentبسازیم و مقایسه کنیم؟
- ✓ چگونه توالی هایی که هم تراز نشده اند را با یکدیگر مقایسه کنیم؟

.....

"جالب اینست: فردی که ۲ عدد ساعت دارد، هیچ گاه زمان دقیق را نمی داند و فردی که یک عدد ساعت دارد، فکر می کند که تنها اوست که زمان را می داند!!!!"

در این بخش چگونگی مقایسه ۲ توالی پروتیینی و یا نوکلئوتیدی را به شما نشان خواهیم داد. هنگامی که نیازمند تعیین خصوصیت هر نوکلئوتید و یا آمینواسید در توالی خود هستید این اطلاعات می تواند برایتان مفید باشد.

برای مقایسه چند توالی، ابتدایی ترین نیاز، انتخاب چندین توالی برای مقایسه و انتخاب هم ترازها می باشد. در این فصل چگونگی برآوردن این نیاز را به شما آموزش خواهیم داد. به عبارت دیگر به شما نشان خواهیم داد که چگونه از میان توالی های گوناگون، چندین توالی را به عنوان یک Multiple sequence alignment انتخاب کنید. ۳ روشی که توانایی پاسخ یافتن برای بیشتر نیازهای شما را دارند عبارتند از: ClustalW ،که قابل استفاده برای همه است، Muscle ، که بسیار سریع است، و Tcoffee، که ابزار بسیار دقیقی است که به شما امکان ترکیب کردن توالی ها و ساختارهای گوناگون را می دهد.

همیشه توجه داشته باشید که ایجاد توالی های به ظاهر خوب ولی در عین حال نامناسب کار بسیار آسانی است. ما در این فصل به شما نحوه ی ارزیابی Alignment های مورد نظر و انتخاب مناسب ترین گزینه ها را آموزش خواهیم داد.

گاهی Multiple sequence alignment در دست شما چندان قابل استفاده و مناسب نخواهند بود. در این فصل خواهید دید،در مواجه با این مشکل چگونه برای این توالی ها جایگزین بیابید و همچنین به شما خواهیم آموخت که چگونه قبل از هم تراز کردن توالیهایتان آن ها را با هم مقایسه کنید. برای آشنایی با این دو موضوع، ۲ راه به شما نشان داده خواهد شد: Gibbs sampler: برای تشخیص قطعات مشابه با طول یکسان و Pratt: یک نوآوری برای کشف الگوی PROSITE.

یک multiple alignment مؤثرترین وسیله تحقیقاتی می باشد که شما می توانید در بیوانفورماتیک تولید کنید.

ما معمولا به موقعیت هایی توجه می کنیم که در کار ما استفاده ای ندارند! اگر شما از اینکه multiple alignment کمکی به شما خواهد کرد و یا نه، مطمئن نیستید و یا به عبارتی دیگر مطمئن نیستید که چگونه از آن استفاده کنید، حتما ادامه ی بخش را بخوانید.

این نکته را به یاد داشته باشد که multiple alignment می تواند شما را در پیش بینی ساختارهای پروتیینی به خصوص پیش بینی عملکرد پروتیین ها یاری دهد.

با این وجود ساخت یک multiple alignment بر پایه ی یک علم دقیق و کامل نیست. درواقع این امر بیشتر به هنر متکی است تا علم. شما باید بتوانید از همه ی دانسته های خود در مورد بیوانفورماتیک و بیولوژی در آن استفاده کنید. در این فصل با تمام اسرار این فن آشنا خواهید شد، تا بتوانید یک multiple alignment متناسب با نیزتان را بسازید.

اگر شما یک alignment دارید و تنها می خواهید آن را اصلاح کنیدو یا ظاهر آن را برای انتشار عمومی بیارایید بهتر است که به فصل ۱۰ مراجعه کنید. اطلاعات موجود در آن بخش بسیار به شما کمک خواهد کرد.

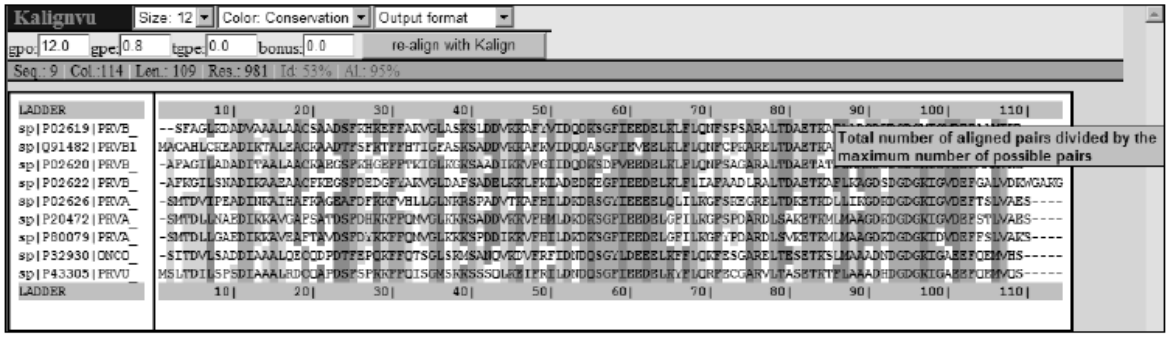
چگونه موقعیت هایی که Multiple sequence alignment کمک کننده نخواهد بود را تشخیص دهیم؟

هنر استفاده از Multiple sequence alignment این است که بدانید در چه مواردی نباید از Multiple sequence alignment استفاده کنید!!! متد های مرتبط با Multiple sequence که در این فصل به شما معرفی خواهد شد چندان در یافتن توالی های مناسب برای یک پروژه ی توالی یابی کارآمد نخواهند بود. اگر توالی های شما شامل مجموعه ای از "توالی های Shotgun" – توالی هایی کوتاه و Overlap- بوده و یا خواستار تغییر فرمت از EST cluster به توالی ژنی هستید، استفاده از Multiple sequence alignment به نفع شما نخواهد بود و استفاده از ClustalW، Tcoffee، Muscle می تواند نتایجی مأیوس کننده و نامتعارف را در بر داشته باشد.

اگر با چنین مشکلی روبرو شدید می توانید از ابزارهایی مانند Phred و phrap و cap3 که برای فرآیند جمع آوری توالی ها اختصاصی هستند، استفاده کنید. البته متأسفانه به علت پیچیدگی این ابزارها، در این کتاب صحبت چندانی از آن ها نشده است و به ذکر اشاره ای در فصل ۵ اکتفا کرده ایم.

مورد دیگری که استفاده از Multiple sequence alignment چندان مفید فایده نخواهد بود، زمانی است که هیچ توالی همولوگی با توالی خود را در پایگاه های داده ای نیابید. اگر با چنین مشکلی برخورد کردید می توان شما را یک انسان بدشانس نامید، چون هیچ راه دیگری برای ساختن یک Multiple sequence alignment نخواهید داشت. البته شاید

بخواهید شانس خود را با استفاده از معیارهای عملکردی و همچنین استفاده از Pratt امتحان کنید، ولی باز هم انتظار معجزه نداشته باشید.



شکل ۱-۹: یک Multiple Sequence Alignment

چگونه می توان با Multiple sequence alignment به سیر تحقیقات خود کمک کرد؟

با در دست داشتن توالی هایی مناسب، ساخت یک Multiple sequence alignment خوب کار چندان سختی نخواهد بود. اگر بخواهیم صادق باشیم، باید بگوییم که ایجاد alignment ها با دست – به جای کامپیوتر- نیز امکان پذیر است. در واقع Alignment های دست ساز مانند غذاهای خانگی می مانند: اگر توسط یک آشپز ماهر پخته شوند، هیچ چیز نمی تواند جای آن را بگیرد. البته واضح است که برای افرادی مثل ما، که در این زمینه چندان مهارت نداریم استفاده از روش های اتوماتیک مناسب تر خواهد بود.

هدف از ساخت multiple alignment ها، قرار دادن آمینواسیدها و نوکلئوتید ها در گروه هایی است که مبنای تشکیل این گروه ها، میزان شباهت آن ها در یک خصوصیت خاص می باشد. در جدول ۱ - ۹ ، چهار معیار اصلی در انتخاب گروه ها نشان داده شده است:

معیار	مفهوم
شباهت ساختاری	آمینواسیدهایی که در هر کدام از ساختار ها نقش یکسانی را دارند، در یک گروه قرار می گیرند. برنامه ی Structure- superposition تنها برنامه ای است که از این معیار استفاده می کند.
شباهت تکاملی	آمینواسیدها و یا نوکلئوتیدهایی که دارای جدی مشترک هستند، در یک گروه قرار می گیرند. هیچ برنامه ی اتوماتیکی واضحا از این معیار استفاده نمی کند، ولی در عین حال تمامی این قبیل برنامه ها، سعی در ارائه ی توالی هایی با این خصوصیت دارند.
شباهت عملکردی	آمینواسیدها و نوکلئوتیدهایی که دارای عملکرد

	یکسانی باشند در یک گروه قرار می گیرند. هیچ برنامه ی اتوماتیکی از این معیار به طور واضح استفاده نمی کند، ولی اگر شما علاقه مند به رعایت این معیار در گروه های خود هستید، برنامه هایی وجود دارند که می توانید اعمال این معیار را از آن ها بخواهید و حتی می توانید چنین معیاری را با تغییرات دستی اعمال کنید.
شباهت توالی	آمینواسیدهایی که در یک گروه قرار می گیرند مربوط به Alignment ی با بیشترین شباهت می باشند. اکثر برنامه های موجود از این معیار در آنالیزهای خود استفاده می کنند، چون از طرفی کار با آن آسان بوده و از طرف دیگر، هر گاه دو توالی به یکدیگر شباهت بالایی داشته باشند(دارای معیار شباهت توالی)، طبیعتا شباهت های ساختاری، عملکردی و تکاملی آن ها نیز مستعجل می باشد.

جدول ۱-۹: معیار های اصلی برای ساخت یک Multiple Sequence Alignment

۳ معیار ابتدایی که در جدول بالا معرفی شدند- برخلاف معیار چهارم- معنی کاملا واضح و روشنی دارند. در عین حال زمانی که ما به توالی هایی با شباهت بالا دسترسی داشته باشیم، به راحتی می توانیم یک Multiple sequence alignment بسازیم که بازتابی از شباهت های ساختاری، عملکری و تکاملی موجود بین توالی های مجموعه می باشد.البته برای انجام چنین کاری طبیعتا نیاز به یک توالی مبدأ برای انتخاب سایر توالی ها داریم. در ادامه ی بخش نحوه ی انتخاب این توالی را به شما نشان خواهیم داد.

در جدول ۲ - ۹ با تعدادی از موارد کاربرد Multiple sequence alignment آشنا خواهید شد.

همیشه به یاد داشته باشید که: علت انتخاب و یا انگیزه تان برای استفاده از Multiple sequence alignment هرچه می خواهد باشد، کاربردی بودن آن در بسیاری از موارد به این علت است که Multiple sequence alignment می تواند یک تصویر آنی از جریان های تشکیل دهنده ی سیر تکاملی را به شما می دهند. وقتی به توالی های پروتیینی و یا DNA نگاه کنید متوجه می شود که قوانین حاکم بر آن ها بسیار سخت و محکمند:

- ✓ آمینواسیدها و یا نوکلئوتیدهای مهم امکان جهش ندارند. برای مثال: جایگاه های فعال آنزیم ها تحت حفاظت بیشتری نسبت به بقیه مناطق هستند.
- ✓ مناطقی با ارزش پایین تر، به راحتی تغییر می کنند؛ گاهی تصادفی و گاهی طبق برنامه(برای تطابق با عملکرد جدید توالی).

شما می توانید از نکته ی بالا این نتیجه گیری را بکنید که: هر گاه به یک مجموعه ای از توالی ها نگاه می کنیم می توانیم فرضیه ای با این مضمون ارائه دهیم: مناطق محافظت شده(مناطقی از توالی ها که دارای آمینواسیدهای یکسانی هستند) دارای اهمیت بیشتری در عملکرد توالی نسبت قسمت های دیگر توالی(قسمت هایی که دارای نسبت به یکدیگر متفاوتند) هستند. هرچند چنین مقایسه ای را می توان بین دو توالی نیز انجام داد ولی استفاده از تعداد توالی بیشتر به ما در تمییز بین مناطق حفاظت شده و نشده کمک بیشتری خواهد کرد.

Application	*Procedure*
Extrapolation	A good multiple alignment can help convince you that an uncharacterized sequence is really a member of a protein family. Alignments that include Swiss-Prot sequences are the most informative. Use the ExPASyBLAST server (at `www.expasy.ch/tools/blast/`) to gather and align them.
Phylogenetic analysis	If you carefully choose the sequences you include in your multiple alignment, you can reconstruct the history of these proteins. Use the Pasteur Phylip server at `bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html`.
Pattern identification	By discovering very conserved positions, you can identify a region that is characteristic of a function (in proteins or in nucleic-acid sequences). Use the logo server for that purpose: `www-lmmb.ncifcrf.gov/~toms/sequencelogo.html`.
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain (PSSM). You can use this profile to scan databases for new members of the family. Use NCBI-BLAST to produce and analyze PSSMs: `www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm`.
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potentially similar binding sites. Use the Gibbs sampler to identify these sites: `bayesweb.wadsworth.org/gibbs/gibbs.html`
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for both proteins and RNA. Sometimes it can also help in the building of a 3-D model.
nsSNP analysis	Various gene alleles often have different amino-acid sequences. Multiple alignments can help you predict whether a Non-Synonymous Single-Nucleotide Polymorphism is likely to be harmful. See the SIFT site for more details: `blocks.fhcrc.org/sift/SIFT.html`.
PCR analysis	A good multiple alignment can help you identify the less-degenerated portions of a protein family, in order to fish out new members by PCR (polymerase chain reaction). If this is what you want to do, you can use the following site: `blocks.fhcrc.org/codehop.html`.

جدول ۲-۹: کاربرد های اصلی Multiple sequence alignment

نکته ای دیگر اینکه Multiple sequence alignment یک راه فوق العاده برای بیان نتیجه ی کارتان است. Multiple sequence alignment به شما این امکان را می دهد که اطلاعات زیادی را در یک شماء قرار داده و همچنین به طور باورنکردنی امکان تعیین تناقضات و مشکلات احتمالی را به شما می دهد. اگر شما می خواهید به نکته ای – بر کاغذ و یا مقابل یک جمعیت- اشاره کنید، یک Multiple sequence alignment خوب، دست ساز و خوش نقش، بهتر از ۳ میلیون توده نوکلئوتیدی، قابل فهم خواهد بود.

با وجود تمام مطالبی که گفته شد، یک Multiple sequence alignment توانایی درمان آسم، آرتریت روماتویید، کمردرد و یا طاسی را ندارد. حتی یک Multiple sequence alignment قرار نیست که شما را قدرتمند سازد و یا اینکه توانایی شما را بالا ببرد. لیکن اگر به دلایل بیشتری برای متقاعد شدن نسبت به منفعت Multiple sequence نیاز دارید، بهترین انتخاب برای شما می تواند این باشد که اصلا آن را نسازید!!!

چگونه مناسب ترین توالی ها را انتخاب کنیم؟

کسانی که تاکنون تجربه ی کار در آزمایشگاه بیولوژی را داشته باشند، حتما به این موضوع پی برده اند که بیولوژی مولکولی بسیار شبیه به هنر آشپزی می باشد!!!! هنر اشپزی تماما در مورد انتخاب عناصر مناسب، نحوه ی صحیح قرار دادن این عناصر در دیگ و در زمان مناسب می باشد. بیولوژی مولکولی نیز چنین است. در ساخت یک Multiple sequence قبل از ساخت alignment مورد نظر باید توالی های مناسب را انتخاب کنیم. توالی هایی را که انتخاب می کنید، همه عضوی از یک خانواده ی پروتیینی بزرگی خواهند بود که دارای اجدادی مشترکند. البته این خانواده به قدری بزرگ می باشد که توانایی انتخاب و گردهمایی همه ی اعضای آن امکان پذیر نیست. بنابراین انتخاب اینکه کدام اعضا در گروه شما قرار بگیند خود یک هنر است. اگر می خواهید در این بازی پیروز باشید باید هدف اصلیتان از انتخاب این توالی ها را بدانید و همچنین با نحوه ی کار با برنامه های مرتبط با Multiple sequence نیز آشنا باشید.

برای آموختن این برنامه ها بهتر است که با آن ها در عمل آشنا شوید و برای آن به مقدماتی همچون "وجود تعدادی توالی" نیاز دارید. این کار مانند مسئله ی قدیمی مرغ/ تخم مرغ است که هیچ راه حل ساده ای برای حل آن وجود ندارد. پیشنهاد ما این است که قسمت بعدی را خوانده و سپس به قسمت های قبلی که توضیح داده شد برگردید. بعد از چند دور کار با برنامه، مسایل برای شما کاملا جا خواهد افتاد.

انواع توالی هایی که شما در پی آن خواهید بود کدامند؟

اکنون فرض می کنیم که شما کار را با توالی های مورد نظر خود آغاز کرده اید. شما می خواهید بر روی این توالی مطالعات کاملی را انجام دهید و برای این منظور باید یک multiple alignment بسازید. جدول ۳ – ۹ مواردی که در هنگام انتخاب این توالی ها باید مورد توجه قرار دهید را به طور خلاصه بیان کرده است:

مشکل	راه حل
وقتی هم پروتیین و هم DNA در اختیار داشته باشیم؟	تا آن جا که امکان پذیر است از فرمت پروتیینی استفاده کنید چون شما به راحتی می توانید در صورت لزوم بعد از ساخت توالی چندگانه ی خود آن را به فرمت DNA مبدل سازید.
وقتی تعداد زیادی توالی در اختیار داریم؟	با ۱۰ – ۱۵ توالی کار را شروع کنید. از کار با بیش از ۵۰ توالی خودداری کنید.
وقتی توالی های بسیار متنوعی در اختیار داریم؟	وجود توالی هایی که کمتر از ۳۰٪ شباهت با بیش از ۵۰٪توالی های دیگر دارند، معمولا مشکل ساز خواهند بود.
وقتی توالی های بسیار مشابه در اختیار داریم؟	چنین توالی هایی چندان سودمند نخواهند بود!!! وقتی توالی هایی در اختیار دارید که بیش از ۹۰٪ شباهت با توالی های دیگر مجموعه دارد – مگر در مواردی که دلیل موجهی دارید– آن ها را از مجموعه حذف کنید.
وقتی توالی هایی ناتمام در اختیار داریم؟	معمولا برای برنامه های مرتبط با Multiple sequence

	alignment ترجیح می دهند که توالی هایی با اندازه ی یکسان را مورد بررسی کنند. وجود توالی های کامل با توالی های کوچک و ناقص در یک مجموعه می تواند کار را مشکل کند.
وقتی توالی هایی با مناطق تکراری داشته باشیم؟	توالی هایی که مناطق تکراری دارند معمولا مشکل زا خواهند بود. در این موارد بهتر است که خود این قسمت ها را با برنامه های Dotlet و Lalign(بخش ۸) جدا کرده و یک multiple alignment مستقل با این قسمت ها بسازید.

جدول ۳-۹: چند نکته در انتخاب توالی ها

بهتر است که از پروتیین استفاده کنیم و یا از DNA؟

اگر توالی که نظر شما را به خود جلب کرده است، یک توالی بدون کد ژنتیکی است، واضح است که هیچ انتخابی به جز DNA ندارید. البته بهتر است این نکته را به یاد داشته باشید که کار با این توالی های DNA می تواند بسیار مشکل باشد. اگر نتوانستید با استفاده از توالی هایی که از شباهت آن ها اطمینان دارید یک alignment مناسب بسازید، می توانید از روش های Local multiple alignment استفاده کنید. از برنامه هایی که در این زمینه می توان نام برد شامل : Gibbs sampler و یا Pratt می باشد(برای یافتن اطلاعاتی بیشر در زمینه ی این دو متد ، بخش " مقایسه توالی هایی که امکان ایجاد هم ترازی بین آن ها نیست" را در همین فصل مطالعه کنید.)

هنگامی که شما از یک توالی پروتیینی برای کار خود استفاده کنید ، بهترین نتیجه را از یک برنامه Multiple sequence alignment به دست خواهید آورد. دلیل آن را هم می توان در این دو نکته خلاصه کرد که اولا یک توالی پروتیینی ۳ مرتبه نسبت به توالی DNA مربوط به خودش کوتاه تر می باشد و ثانیا در یک توالی پروتیینی از زبان ۲۰ آمینواسیدی استفاده می شود که بسیار سودمنتر خواهد بود.

اگر شما بر انجام یک آنالیز فیلوژنیک بر روی توالی های DNA کد گزاری شده مصر هستید، با رعایت نکات زیر به نتایج بهتری دست خواهید یافت:

۱. توالی های DNA خود را به پروتیین ترجمه کنید.
۲. فرامین multiple alignment را بر روی توالی های پروتیینی اعمال کنید.
۳. توالی های DNA خود را با استفاده از برنامه ی pal2nal (coot.embl.de/pal2nal) به توالی های پروتیینی و یا با استفاده از)www.tcoffee.org (به توالی های پروتوژن –اگر توالی DNA اصلی را ندارید- تبدیل کنید.

اگر به علت شباهت ناچیزی که در توالی های پروتیینی شما وجود دارد، امکان ایجاد هم ترازی میان آن ها نیست، اطلاعات مرتبط با DNA آن رشته ها نیز کمک کننده نخواهند بود. این به آن علت است که از طرفی کدهای ژنتیکی چندگانه هستند و از طرف دیگر درحالی که ۲۰ آمینواسید وجود دارد تنها ۴ نوکلئوتید موجود می باشد. اگر در فرمت پروتیینی علامات چندان مورد توجهی مشاهده نکردید، مطمئن باشید که بر روی فرمت DNA مربوطه هم چیزی نخواهید یافت.

چگونه تعداد مناسب توالی را انتخاب کنیم؟

هیچ پاسخ قطعی برای این سوال وجود ندارد. نمی توان گفت که حتما با ۴۲ و یا ۷ توالی شروع کنید. در سال هایی نه چندان دور پاسخ به این سوال آسان بود: اگر توالی های موجود در پایگاه های اطلاعاتی برایتان کافی نیست، هر تعداد توالی

که می توانید بیابید و استفاده کنید!!!. ولی چنین دستورالعملی امروزه با توجه به اندازه پایگاه های اطلاعاتی غیر ممکن به نظر می رسد. چون ممکن است ما صدها و یا حتی هزاران توالی که می توانند برای کار ما مناسب باشند را بیابیم ولی این نمی تواند به این معنی باشد که لزوما از همه ی آن ها استفاده کنیم.

به عقیده ی ما بهتر است که کارتان را با تعدادی اندک توالی- نزدیک ۱۰ تا ۱۵- شروع کنید و اگر به نتایج مورد توجهی رسیدید، کارتان را با تعداد توالی بیشتری ادامه دهید. در بسیاری از موارد دلیلی برای استفاده ی بیش از ۵۰ توالی وجود ندارد، مگر در مواردی که شما تصمیم به ساخت یک درخت فیلوژنیک بزرگ داشته باشید.

اگر کار را با هزاران توالی آغاز کنید بلافاصله به مشکل بر خواهید خورد. دلایل مناسبی بر این مدعا وجود دارند:

- ✓ محاسبه Alignment های بزرگ دشوار خواهد بود: سرورهای عمومی منابع گسترده ای دارند. کار با تعداد توالی بالا در این سرورها بسیار طول می کشد(البته اگر انجام شود!!!) و می تواند عاملی باشد که شما را از امتحان پارامترهای دیگر و انتخاب گزینه های متفاوت بر حذر دارد.
- ✓ ساخت alignment های بزرگ سخت می باشد: برنامه های Multiple alignment از عهده ی اداره ی مجموعه های بسیار بزرگ بر نمی آیند.
- ✓ نمایش Alignment های بزرگ دشوار است: امکان پرینت از همه ی آن ها وجود ندارد و اگر بخواهید آن را در کامپیوتر خود ببینید، به علت حجم بالا، بسیار کند خواهد بود و همچنین اگر ستون های نمایش داده ها از یک صفحه بزرگتر شود، امکان تفسیر آن ها وجود نخواهد داشت.
- ✓ انجام هم ردیفی های بزرگ دشوار است: برنامه های "ساخت شجره نامه" و " پیش بینی کننده ساختاری" به راحتی نمی توانند با آن ها کار کنند. تصحیح alignment های بزرگ دشوار است: طبیعتا شما ترجیح می دهید که یک Multiple alignment قابل اطمینان بسازید و مقدمه ی ساخت چنین مجموعه ای اطمینان از درستی اعضای انتخاب شده آن است. در واقع شما باید از خویشاوندی اعضای گروه خود اطمینان داشته باشید. یکی از دلایلی که می تواند باعث نگرانی در این زمینه باشد این است که برنامه های Multiple alignment امکان اشتباه در این زمینه را دارند. و مشکل بزرگتر این است که این خطاها با یکدیگر جمع نمی شوند بلکه ضرب خواهند شد، به عبارتی رشد تصاعدی خواهند داشت. نکته ی بدیهی دیگری که وجود دارد این است که هرچه تعداد توالی ها بیشتر باشد امکان بروز این خطا و عواقب شدیدتر آن بیشتر خواهد شد. بنابراین ترجیحا باید کار را با تعداد اندکی توالی آغاز کرد و به مرور و کم کم به تعداد توالی ها اضافه کرد تا جاییکه تمام توالی های مورد نظر را به مجموعه وارد کنیم.

اکنون که دیگر می دانید چگونه تعداد توالی مناسب برای شروع کار خود را تعیین کنید به سوال آخری که باید حواب دهید می رسیم. یعنی در آخر باید از خود سوال کنید که چه نوعی و چه میزانی از شباهت را مورد توجه قرار دهیم؟ باید توالی هایی با شباهت زیاد را انتخاب کنیم و یا برعکس، توالی هایی با شباهت اندک بیشتر مورد توجه هستند؟

ترکیب مناسب شباهت و اطلاعات جدید

اگر فکر می کنید توالی هایی که بیشترین شباهت را به هم دارند بهترین alignment ها را ایجاد می کنند، باید بگوییم که کاملا درست فکر می کنید. هرچند بیان این نکته ضروری است که صرف ساخت یک Multiple sequence alignment کافی نیست بلکه یک Multiple sequence alignment باید قابل استفاده و مفید هم باشد.

برای مثال alignment ی را در نظر بگیرید که شامل توالی های بسیار مشابه است. چنین alignment ی اطلاعات زیادی را در اختیار شما قرار نخواهد داد. از چنین alignmentی برای حدس یکسری داده ها می توانید استفاده کنید، ولی برای یک پژوهش فیلوژنیک، پیش بینی ساختاری و یا هر کدام از موارد موجود در جدول۳-۹ قابل استفاده نخواهد بود. برای رسیدن به چنین نتیجه هایی باید توانایی مشاهده ی توام جهش های موجود را در توالی ها داشته باشد، در حالی که در چنین alignmentی که تمام قسمت های دست نخورده و فاقد جهش موجود است نمی توان به چنین نتایجی دست یافت.

توالی های بسیار دور از نظر شباهت نیز خیلی قابل استفاده نخواهند بود. در واقع دو دسته توالی هایی که چندان مورد علاقه ی برنامه های Multiple sequence alignment نیستند شامل:

- ✓ توالی های بسیار متفاوت از اکثریت توالی های موجود در گروه.
- ✓ توالی هایی که برای تبدیل شدن به یک توالی همتراز نیاز به حذف و یا اضافه ی قسمت بزرگی از ساختارشان وجود دارد.

انتخاب توالی برای یک Multiple alignment مانند هنر تشکیل یک گروه می باشد. برای تشکیل یک گروه ما به افرادی نیاز داریم که هر کدام نسبت به دیگری یک خصوصیت خاص داشته باشند ولی در عین حال آن قدر از هم فاصله نداشته باشند که امکان تطابقشان نباشد.

قانون کلی در انتخاب توالی ها این است که تا حد امکان نباید توالی ها از هم فاصله ی زیادی داشته باشند و در عین حال برای تبدیل شدن به یک هم تراز مناسب هم نیاز به تغییر چندانی در آن ها نباشد. پیدا کردن توالی هایی که با این دو معیار نسبتا متناقض هماهنگ باشند نیاز به استراتژی های خاصی دارد. مراحلی که در زیر بیان شده اند، خط مشیی کلی به شما نشان خواهند داد که باید هنگام انتخاب توالی های خود آن ها را در نظر داشته باشید:

۱. با تعداد کمی توالی کار را شروع کنید: برای توضیح بیشتر به قسمت" جمع آوری توالی ها با استفاده از Online BLAST server" که در ادامه ی بخش با آن روبه رو می شوید مراجعه کنید.
۲. Multiple alignment را با یکی از سرورهای معرفی شده در همین بخش مورد ارزیابی قرار دهید: برای توضیح بیشتر به قسمت " چگونه مناسب ترین روش را برای Multiple sequence alignment انتخاب کنیم؟ " مراجعه کنید.
۳. کیفیت alignment را از نظر ظاهر ارزیابی کنید: به قسمت " چگونه Multiple sequence alignment را تفسیر کنیم؟" مراجعه کنید.
۴. اگر ظاهرا alignment شما مناسب به نظر می آید، توالی های موجود را نگه دارید: یک alignment مناسب alignmentی است که دارای مناطقی دست نخورده و محافظت شده باشد. اگر در alignment شما چنین توالی هایی موجود است، می توانید از مناسب بودن آن مطمئن باشید و با اضافه کردن توالی هایی جدید آن را گسترش دهید.
۵. اگر تفسیر alignment تان مشکل است:
   a. توالی ها را با دقت بیشتری مورد بررسی قرار دهید. سعی کنید توالی های مشکل ساز را که اکثرا توالی هایی با شباهت پایین و یا توالی هایی با مناطق حذف و اضافه ی طولانی هستند را از گروه خارج سازید.
   b. هم تراز کردن را با مجموعه ای کوچکتر دوباره تکرار کنید.
   c. تا زمانی که به یک Multiple sequence alignment با قابلیت تفسیر آسان دست یابید، مراحل گذشته را تکرار کنید.

اگر مطمئن شدید که هر توالی ۳۰ تا ۷۰٪ با نیمی از توالی های مجموعه شباهت دارد، در واقع می توان گفت تعادلی منطقی میان داده های جدید و alignment شما برقرار کرده اید.

شما می توانید با استفاده از ابزار معرفی شده در فصل ۶ و مقایسه ی دو به دو توالی ها، قبل از اضافه کردن هر توالی به مجموعه، از درستی انتخاب خود اطمینان حاصل کنید. البته ما چنین کاری را چندان به شما توصیه نمیکنیم. چون علی رغم دقت آن، بسیار وقت گیر خواهد بود.

چگونه توالیهایمان را باروشی صحیح نام گذاری کنیم؟

شاید نامگذاری توالی ها یک مسئله ی جزئی و کم اهمیت به نظر آید، ولی اینگونه نیست. برنامه های Multiple sequence alignmentهیچ روش استانداردی برای نامگذاری توالی ها ندارند. با رعایت ۴ قانون ذکر شده در زیر، هرگز به مشکل برنخواهید خورد:

- ✓ هرگز در نامگذاری توالی های خود از فضای خالی(white space) استفاده نکنید.
- ✓ از علایم خاص استفاده نکنید. فضاهای خالی را با حروف، اعداد و خط فاصله پر کنید. از هرگونه علایم به خصوص علایمی که مورد توجه توالی های خاص هستند نظیر @، #، _ ، ^ و .... استفاده نکنید.
- ✓ هرگز نام های طولانی تر از ۱۵ حرف را انتخاب نکنید.
- ✓ هرگز به دو توالی متفاوت در مجموعه ی خود یک نام را ندهید. هرچند بعضی از برنامه ها آن را خواهند پذیرفت ولی برنامه هایی مانند ClustalW چنین نخواهند بود.

اگر از این قوانین استفاده نکنید، بعضی از برنامه های Multiple sequence alignment نام توالی های شما را بدون اطلاعتان تغییر خواهند داد.

چگونه توالی های مورد نیازمان را با استفاده از سرورهای آنلاین BLAST جمع آوری کنیم؟

۲ نوع توالی که احتمالا شما آن ها برای اضافه کردن به multiple alignment خود مورد توجه قرار خواهید داد عبارتند از:

- ✓ توالی های مشخص شده (Characterized sequences): این توالی ها توالی هایی هستند که اطلاعات و تفاسیر مناسب و خوبی برایشان وجود دارد. قطعا هر محققی مایل است که چنین توالی هایی، در alignment او قرار بگیرد، چون این توالی ها اطلاعات بیولوژیک خوبی را به همراه دارند و همچنین استفاده از آن ها این امکان را به ما می دهد که خصوصیت های مورد نظرمان را که در آن ها وجود دارد مورد استفاده قرار دهیم.
- ✓ توالی های غیر مشخص شده (Uncharacterized sequences): این دسته از توالی ها نیز می توانند علاوه بر توالی های موجود در پایگاه داده ها مورد استفاده قرار گیرند. این توالی ها باید اعضای یک خانواده ی بزرگ باشند. هرف اصلی از انتخاب آن ها، تمایز قائل شدن بین موقعیت های دست نخورده و حفاظت شده با دیگر ستون های کم اهمیت می باشد. این توالی ها شما را در به دست آوردن بعضی از تناقضات یاری خواهند داد.

در این بخش، نحوه ی جمع آوری این توالی ها با استفاده از برنامه های BLAST را به شما نشان خواهیم داد. اگر به دانستن اطلاعات کامل تری در این زمینه مشتاق هستید، بهتر است که فصل ۷ را باری دیگر مطالعه کنید، چون در این قسمت حداقل اطلاعات لازم در اختیار شما قرار خواهد گرفت.

با استفاده از BLAST می توانید توالی های همولوگ با یک توالی معین را پیدا کنید. توالی مبدأ می تواند هر توالی که مورد نظر شما هست، انتخاب شود، فرقی نمی کند که پروتیین باشد و یا DNA. هدف اصلی از استفاده از BLAST، پیدا کردن توالی هایی با بیشترین شباهت به توالی مورد نظر می باشد. چنین توالی های مشابهی احتمالا همولوگ می باشند و ما چنین توالی هایی را Matches و یا Hits می نامیم.

انواع متنوعی از سرورهای ارائه دهنده ی سرویس BLAST وجود دارند و شما در انتخاب و استفاده از هر کدام از آن ها مختارید. هرچند لازم به یادآوری است که همه ی این سرورها به یک اندازه برای کار شما مفید نخواهند بود. در جدول ۴ -۹ با ۳ سرور ارائه دهنده ی BLAST که داده هایی با فرمت FASTA (فرمتی مناسب برای انتقال از یک برنامه به برنامه ای دیگر و یا برای انتقال داده های مورد نظر به یک برنامه ی Multiple sequence alignment) را در اختیار شما قرار می دهند، آشنا می شوید.

Address	*What You Can Do There*
www.expasy.ch/tools/blast/	Extract entire sequences, export sequences in FASTA, submit sequences to ClustalW, Tcoffee or MAFFT. Turn the list of Hits into a non-redundant collection of sequences.
npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_blast.html	Extract entire sequences, extract sequences fragments, export sequences in FASTA, submit sequences to ClustalW.
srs.ebi.ac.uk	Submit sequences to ClustalW.

جدول ۴-۹: سرور های BLAST که روش های Multiple-Alignment را با هم ترکیب می کنند

در مراحل ذکر شده، با نحوه ی کار با سرور ExPASy آشنا خواهید شد. هرچند سرور SRS در بعضی از موارد مفید می باشد ولی استفاده از آن نسبت به دو سرور دیگر معرفی شده مقداری سخت تر است.

چگونه توالی های مورد نیاز خود را با استفاده از سرور ExPASy انتخاب کنیم؟

سوال مورد نظر ما که با انجام مراحل زیر به جواب آن خواهیم رسید عبارت است از: " توالی های مناسب برای ساخت یک Multiple sequence alignment مرتبط با پروتیین کیناز وابسته به کلسیم کدامند؟"

با استفاده از این سرور تنها توالی های پروتیینی را به دست خواهید آورد. اگر به دنبال توالی های DNA هستید، از سرور SRS(SRS.ebi.ac.uk) استفاده کنید.

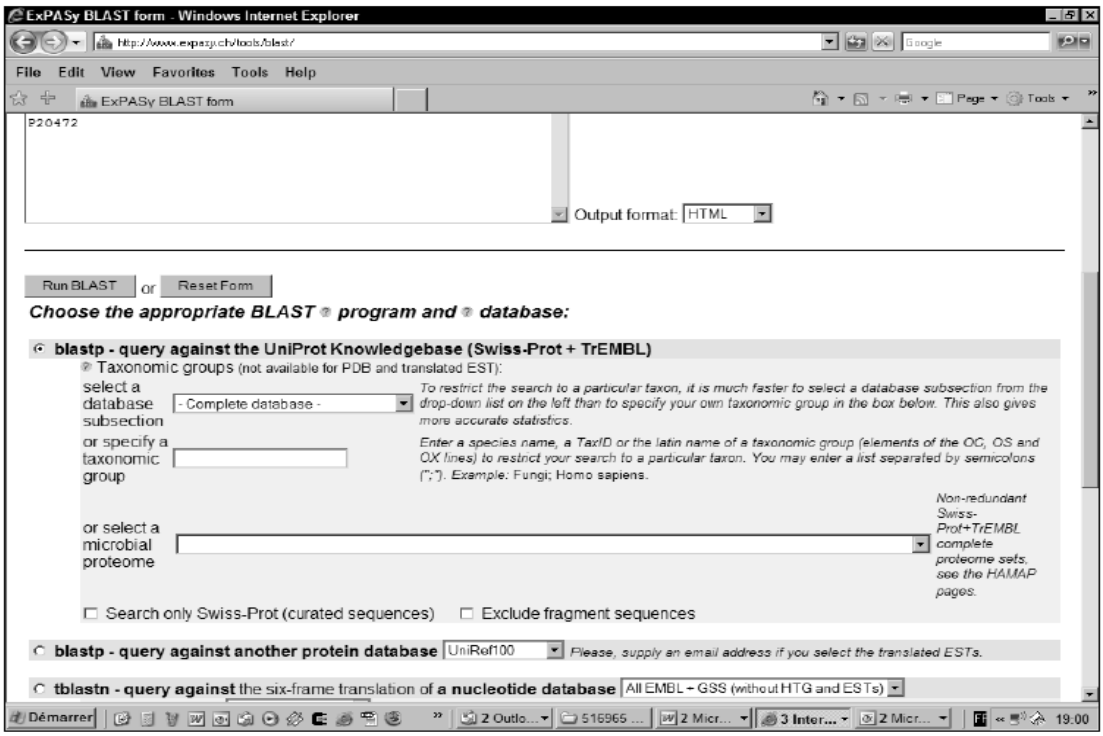
سرورBLAST موجود در سایت PBIL بسیار مشابه با ExPASy است. اگر پایگاه داده ای مورد نظر خود را در ExPASy نیافتید، شانس خود را در PBIL امتحان کنید.

۱. وارد سایت www. Expasy. Ch/tools/blast/ شوید. صفحه ی BLAST مرتبط با ExPASy باز خواهد شد.

۲. شماره مسلسل توالی را در محل تعیین شده وارد کنید (شکل 9-2).

عدد مورد نظر ما در این جا p20472 می باشد که مربوط به " پاروالبومین[1] انسانی" است.

اگر مایلید می توانید توالی مورد نظر را با همان فرمت خام خود وارد کنید. خود برنامه به طور خودکار، فضاهای خالی و اعداد را نادیده خواهد گرفت.



شکل ۲-۹: استفاده از سرور BLAST ExPASy برای جمع آوری توالی

۳. نوع BLAST مورد نظر خود را انتخاب کنید.

اگر در قسمت ۲ ، یک توالی پروتیینی وارد کرده اید، BLASTp را انتخاب کنید. و اگر یک توالی DNA وارد کرده اید، tBLASTn را انتخاب کنید.

۴. در منوی موجود، همان گزینه پیش فرض (complete database) را انتخاب کنید.

این پیش فرض امکان جستجو در پایگاه های اطلاعاتی Swiss-Prot + TrEMBL+ TrEMBL_NEW را به طور همزمان فراهم می کند. اگر نتایج به دست آمده از این جستجو شامل تعداد زیادی از توالی های مشابه با توالی شما بود می توانید با انتخاب یکی از پایگاه های داده ای از همین منو، میزان داده های دریافتی را کاهش دهید.

به یاد داشته باشید که از این منو هر گز گزینه ی translated EST را برای ساخت یک multiple sequence alignment استفاده نکنید. زیرا این گزینه پایگاه داده ای شامل توالی های ناکامل پروتیینی است که ممکن است در روند کار شما اخلال ایجاد کنند.

۵. در پایین صفحه به بخش Options بروید و در قسمت Number of Best Scoring Sequences to Show عدد ۱۰۰۰ را انتخاب کنید.

با انجام این مرحله، امکان یافتن توالی هایی مناسب تر افزایش می یابد.

[1] Parvalbumin

۶. در همین قسمت، در گزینه Number of Best Alignments to Show نیز عدد ۱۰۰۰ را انتخاب کنید.
این مرحله، امکان اظهار نظر در مورد alignment ، قبل از اضافه شدن توالی را امکان پذیر می سازد.

۷. گزینه ی Run BLAST را انتخاب کنید.
بعد از مدت کوتاهی صفحه ی نتایج باز خواهد شد.

۸. در صفحه به دنبال توالی مورد نظر خود بگردید.
با انتخاب هر توالی مورد نظر خود، مربع کنار توالی را انتخاب کنید.

این مرحله، حساس ترین مرحله ی کار می باشد. هیچ قانون مشخصی برای انتخاب توالی های مورد نظر وجود ندارد اما راهنمایی های زیر می تواند راهنمای خوبی برای شما در انتخاب توالی هایتان باشد:

- ✓ بالاترین توالی یافته شده را انتخاب کنید. این توالی معمولا توالی مناسبی است. اگر توالی مناسب شما در اولین رده ها یافت نشد، حتما بعدا آن را به این جا منتقل کنید.
- ✓ در مرحله ی اول آنالیز داده ها، بهتر است که تعداد ۱۰ و یا کمتر توالی انتخاب کنید.
میزان ایده آل، توالی هایی هستند که در محدوده ی ماکزیمم(10^{-40}) و مینیمم(10^{-5}) بهترین عدد E قرار گرفته باشند. (شکل ۳ -۹)
برای اینکه در این مرحله به هدف مطلوب دست یابید، ده تای اول از توالی ها را انتخاب کنید.
- ✓ قبل از اینکه توالی را انتخاب کنید، از مناسب بودن آن اطمینان یابید.

قسمت مرتبط با alignment در انتهای صفحه موجود می باشد. به خصوص به مواردی که عدد E بالاتر از 10^{-10} دارند توجه کنید. زیرا این توالی ها معمولا تناسب مناسبی با توالی مورد نظر ما برقرار می کنند.

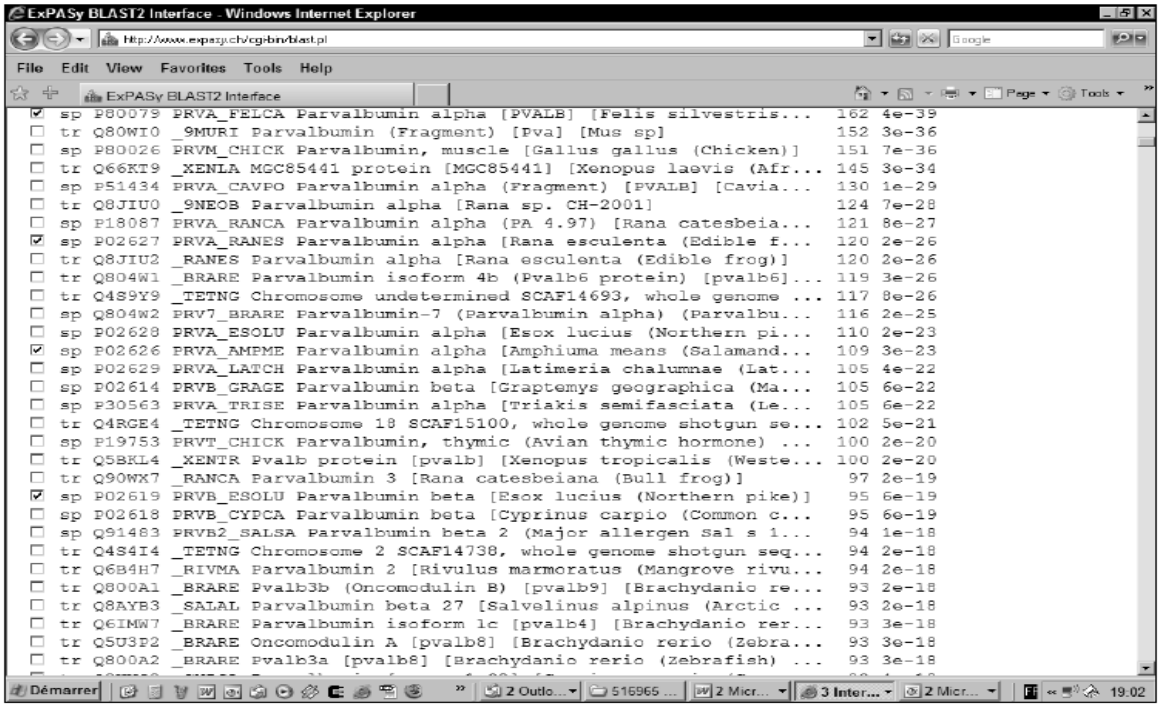
۹. از منوی Send Selected Sequences، فرمت مورد نظر خود را برای نمایش توالی های حاصل انتخاب کنید (شکل ۴-۹).
برای این منظور فرمت های مختلفی وجود دارد:

- ✓ FASTA: با انتخاب این گزینه، فرمت داده های خروجی شما FASTA خواهد بود. با انتخاب file⇨save as option امکان ذخیره داده ها وجود دارد و اگر تمایل داشتید می توانید با copy, past آن در سروری دیگر(برای مثال MUSCLE در www.drives.com) ، از این داده ها در سروری جدید استفاده کنید.
- ✓ ClustalW, Tcoffee, MAFFT: این برنامه ها، برنامه هایی در ارتباط با Multiple sequence alignment می باشند. از هر کدام از آن ها می توانید برای هم تراز کردن توالیهایتان استفاده کنید.

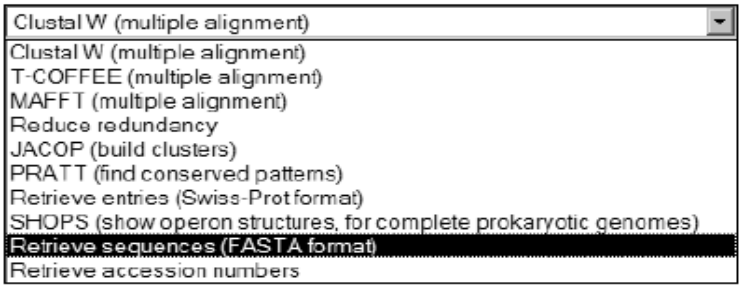
- ✓ Reduce Redundancy: این برنامه بهترین توالی ها را برایتان انتخاب می کنند. اگر تعداد زیادی توالی در اختیار دارید و نمی دانید که کدام را باید انتخاب کنید، استفاده از این برنامه می تواند کمک یار شما باشد.
- ✓ Pratt: با استفاده از این برنامه می توانید مناطق محافظت شده ی موجود در توالی های خود را بدون اینکه align شوند بیابید.

چگونه توالی هایی معین را از Swiss-Prot بیابیم؟

اگر نام و یا شماره ی سریال توالی های مورد نظر خود را می دانید و اگر این توالی ها در پایگاه های داده ای Swiss-Prot و یا TrEMBL موجود باشند، می توانید با استفاده از یکسری امکانات آنلاین ExPASy ، به طور مستقیم آن ها را به مجموعه ی خود اضافه کنید.



شکل ۳-۹: انتخاب توالی ها در خروجی BLAST



شکل ۴-۹: انتخاب گر خروجی در صفحه ExPASy BLAST

اگر از انواعی از سرورBLAST استفاده می کنید که چنین امکاناتی را ندارد، می توانید با کپی گردن شماره مسلسل توالی مورد نظرتان، آن توالی را به صورت دستی به گروه خود اضافه کنید.

به یاد داشته باشید مراحلی که در زیر آموزش داده می شوند تنها زمانی قابل استفاده خواهند بود که توالی مورد نیاز شما در پایگاه داده ای Swiss-Prot و یا TrMBL موجود باشد:

۱. وارد آدرس اینترنتی www.expasy.ch/sprot/sprot-ratrieve-list.html شوید.
۲. قسمت نام فایل را سفید باقی بگذارید تا توالی ها مستقیما برگردانده شود.
۳. در قسمت Format line ، فرمت FASTA را انتخاب کنید.
۴. شماره ی مسلسل توالی خود را در قسمت مربوطه وارد کنید(شکل۵-۹).

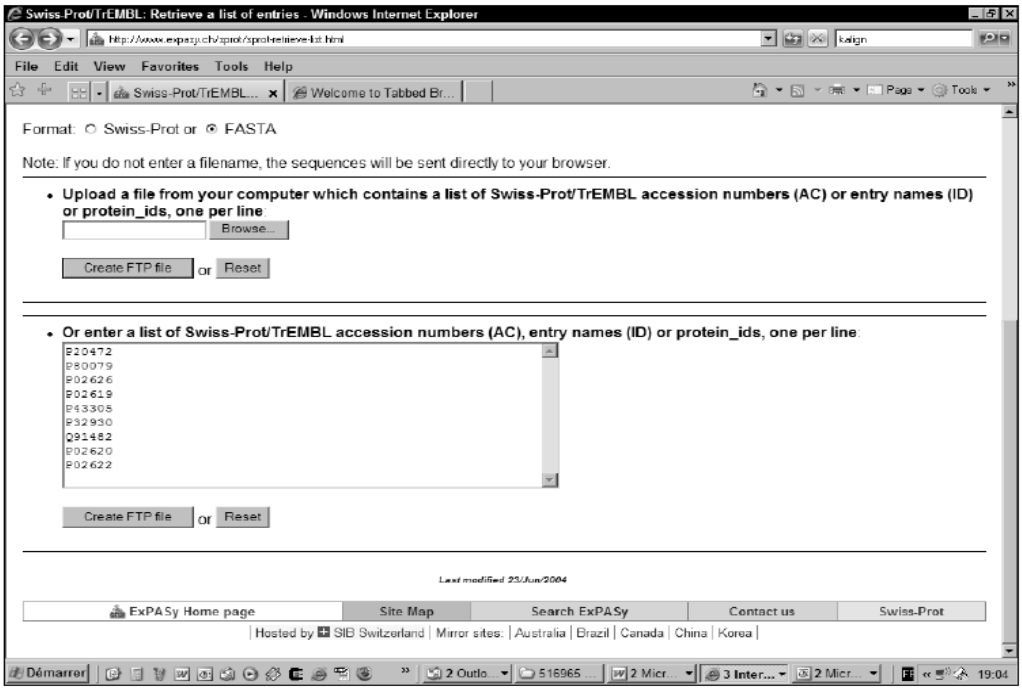
برای هر خط یک شماره مسلسل (یا نام توالی) وارد کنید. برای مثالی که شما دنبال می کنید، کدهای P20472، P80079، P02626، P02619، P43305، P32930، Q91482، P02620 و P02622 را وارد کنید.

اگر کد توالی ها را به صورت ذخیره شده در یک جا داشته باشید می توانید با استفاده از قسمت Upload a File آن ها را به راحتی بار گذاری کنید. شما می توانید این فایل را با برنامه Microsoft Word درست کنید. البته به یاد داشته باشید که این فایل باید با فرمت text باشد.

۵. پایین تر از فیلدی که در آن شماره توالی را وارد کردید، روی گزینه ی Create FTP File کلیک کنید. با انتخاب این گزینه، توالی مورد نظر شما به سرور ExPASy ارائه خواهد شد.

۶. نتیجه را به صورت فایل متنی با انتخاب save⇨file ذخیره کنید.

داده های خروجی این سرور شامل توالی هایی است که نامی بلندتر از ۱۵ حرف دارند. چنین توالی هایی می تواند در سرورهای دیگر مشکل ساز باشند.



شکل ۵-۹: جمع آوری چندین توالی از سرور ExPASy

چگونه روشی مناسب برای Multiple sequence alignment انتخاب کنیم؟

قبل از اینکه شروع به ساخت Multiple sequence alignment کنید، بهتر است این نکته را بدانید که هیچ کدام از روش های موجود بی عیب و نقص نیستند. همه ی آن ها از تخمین و تقریب استفاده می کنند.

برای رسیدن به یک Multiple alignment که بتواند ما را در رسیدن به یک کشف جدید کمک کند، نیاز به ممارست و تمرین زیاد داریم. استراتژی معمولی که وجود دارد شامل "مقایسه نتایج گوناگون به دست آمده با یکدیگر" می باشد.

در این بخش به شما نحوه ی استفاده از ClustalW – رایج ترین بسته ی استفاده شده در Multiple alignment – را نشان می دهیم. این ابزار، ابزاری ساده و بی عیب می باشد. همچنین روش استفاده از t-coffee – یکی از جدید ترین بسته های قابل استفاده در Multiple alignment– را به شما نشان می دهیم. با استفاده از این برنامه می توانید توالی و ساختارها را ترکیب کنید، یک alignment را ارزیابی کنید و یا چندین multiple alignment را با هم ادغام کنید، تا به یک نتیجه ی منفرد دست یابید، و در نهایت شما را با muscle – یکی از سریع ترین روش ها ی alignment– آشنا می کنیم.

چگونه از ClustalW استفاده کنیم؟

این برنامه یکی از رایج ترین برنامه ها برای ساخت Multiple sequence alignment می باشد، به طوری که اگر در یک نشریه ی علمی با multiple alignment برخورد کردید، تقریبا می توانید مطمئن باشد که با کمک ClustalW ساخته شده است.

ClustalW از یک روش جدید و پیشرفته به جای align کردن توالی ها استفاده می کند. این برنامه توالی ها را یک به یک اضافه می کند. فهمیدن اصلی که در زیر به آن اشاره شده است می تواند به شما کمک کند تا بهترین نتیجه را از ClustalW به دست آورید.

اکثریت سرور های ClustalW نتیجه ی یکسانی را در اختیار شما قرار می دهند ولی میانجی های آن ها این امکان را به شما می دهد که انتخاب های متفاوتی پیش رو داشته باشد. در آخر این فصل، لیستی از سرور های ارائه دهنده ی ClustalW به شما ارائه خواهد شد. در پیدا کردن سروری از ClustalW که هم سریع و هم قابل اطمینان باشد باید وقت زیادی صرف کنید.

چگونه از EBI ClutalW استفاده کنیم؟

ClustalW یک نمونه ی کاملا تیپیک از برنامه هایی است که با رویکرد تنظیمات خود نتایجی را به شما عرضه می کنند. تنها زمانی که به دنبال نتایج متفاوتی هستید، باید نگران تنظیمات باشید!!!

قبل از استفاده از برنامه ی ClustalW باید یکسری کارهای مقدماتی را انجام دهید.

مناسب ترین و راحت ترین راه برای استفاده از ClustalW این است که داده های خود را به صورت فرمت FASTA در اختیار سرور قرار دهید. البته این به آن معنی نیست که امکان استفاده از فرمت های دیگری مانند Swiss-Prot و یا PTR و ... را ندارید!!!

اگر می خواهید از ClustalW نتیجه مناسب به دست آورید، دو نکته ی زیر را رعایت کنید:

- ✓ اگر دسته ای از توالی های align شده را به این سرور بدهید، برنامه اختلاف های موجود را حذف نخواهد کرد. این به آن معنی است که اطلاعات ورودی که شما به برنامه می دهید در alignment تهیه شده توسط برنامه تأثیر خواهد داشت.
- ✓ گاهی ترتیب دادن توالی ها به برنامه بر alignment ساخته شده تأثیر خواهد گذاشت. اگر شما این ترتیب را تغییر دهید- حتی با وجود توالی های یکسان- alignment ساخته شده می تواند متفاوت باشد.

اکنون که توالی های مورد نظرتان را در اختیار دارید می توانید کار با ClustalW را شروع کنید:

۱. وارد سایت www.ebi.ac.uk/clastaw شوید.
۲. توالی های جمع آوری شده را در پنجره Sequence وارد کنید.
۳. گزینه FASTA را از قسمت Alignment انتخاب کنید.
۴. از منوی Output Format فرمت مورد نظر را انتخاب کنید.

در این قسمت انتخاب های زیادی وجود دارد. بهترین کار انتخاب ALN – فرمت پیشنهادی خود برنامه- می باشد.

شما این امکان را دارید که هر زمان که مایلید فرمت مورد نظر خود را تغییر دهید. اگر از نوع فرمت انتخابیتان ناراضی هستید، همه ی کار را از ابتدا تکرار نکنید. بهترین کار استفاده از برنامه های Reformat مانند Fmtseq می باشد.

۵. گزینه Input from the Output Order را انتخاب کنید (شکل ۶-۹).

در این جا چند نکته در مورد تنظیم output را به شما یادآوری می کنیم:

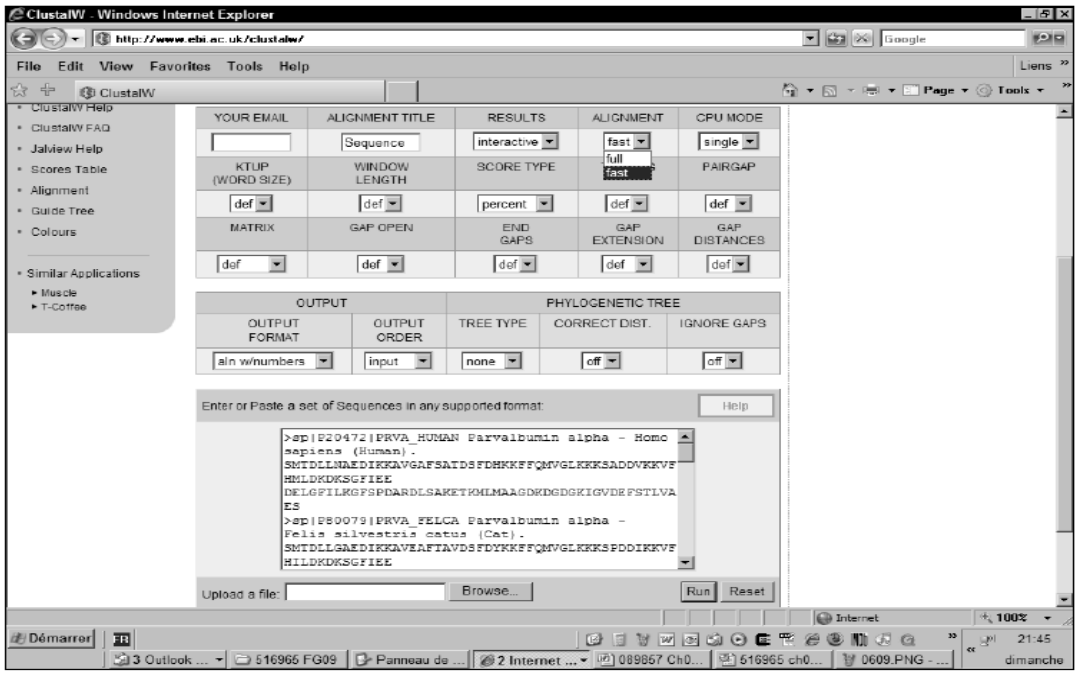
- ✓ را انتخاب می کنید، ClastalW داده های خود را با ترتیب اولیه خود می دهد.
- ✓ وقتی شما در قسمت output order to input گزینه Aligned را انتخاب کنید، توالی ها با ترتیبی که هم تراز شده بودند، ظاهر می شوند.
- ✓ اگر چه ترتیب Aligned اطلاعات بیشتری در اختیار شما قرار می دهد، اما با این فرمت مقایسه alignment های حاصل از روش های مختلف یا با رعایت پارامتر های متفاوت مشکل خواهد شد.

۶. نوع درخت را انتخاب نکنید.

در اینجا نیازی به انتخاب نوع خاصی از درخت نیست. اگر نوعی درخت در اینجا انتخاب کنید ممکن است که درخت بر اساس توالی های align نشده ساخته شود. اگر می خواهید درختی بسازید، ابتدا توالی ها را تنظیم کنید و پس از آن یک درخت فیلوژنیک بر اساس توالی های محاسبه شده بسازید.

۷. گزینه ی Run را انتخاب کنید.

در ابتدا صفحه ای نمایان می شود که صفحه ای واسط است. باید صبر کنید تا صفحه ی نتایج نمایان شود.



شکل ۶-۹: سرور EBI ClustalW

۸. نتایج به دست آمده را ذخیره کنید.

نتایج در ۳ بخش خواهند بود:

- ✓ امتیازات دو به دو (Pairwise scores): این قسمت، قسمت پیش فرض برنامه ی ClustalW می باشد. شما با اطمینان می توانید مندرجات این بخش را نادیده گرفته و حذف کنید.
- ✓ Multiple alignment: این بخش در میانه ی صفحه ی نتایج شما می باشد و متشکل از توالی های شماست. اگر منطقه ی hyperlink که در ادامه ی alignment امده است را کلیک کنید می توانید alignment مورد نظر را نمایش داده و یا با فرمت متنی ذخیره کنید.

اکثر سرور های ClustalW (از جمله EBI) یک درخت راهنما در اختیار شما قرار می دهند (به صورت یک فایل dnd.). این درخت، یک درخت فیلوژنتیک نیست. برای به دست آوردن درخت فیلوژنتیک از یک سرور ClustalW، لازم است از یک multiple sequence alignment واقعی استفاده شود، نه یک مجموعه توالی های هم تراز نشده (در فصل ۱۳ در این مورد بیشتر بحث می شود).

چگونه می توانیم پارامترهای ClustalW را تغییر دهیم؟

۳ پارامتر در این برنامه، تأثیر زیادی بر نتیجه ی کار شما خواهند داشت: substitution matrices, شکاف opening-penalties, شکاف extension penalties-. جدول ۵ -۹ این پارامترها و تغییرات احتمالی بر نتیجه ی کارتان را پس از تغییر میزانشان به شما نشان داده است.

پارامترها	اثرات
substitution matrices	با تغییر این پارامتر می توان میزان جهشی که می خواهیم در نظر گرفته شود را تعیین کنیم. تخمین زدن میزان اثر تغییر این پارامتر کار آسانی نیست و میزان خاصی به عنوان بهترین عدد وجود ندارد. اگر توالی های شما بسیار مشابه باشند، تغییر این پارامتر چندان تأثیری نخواهد داشت ولی اگر میزان شباهت توالی های شما کم باشد تغییر پارامتر می تواند تغییرات زیادی را بر نتیجه ی شما بگذارد.
شکاف opening penalties-	با تغییر این پارامتر میزان شکاف ها را می توانیم تحت کنترل گیریم. در انتخاب بالاترین حد آن، تقریبا هیچ شکافی نمیتوان در توالی وارد کرد. تغییر در این پارامتر تأثیر چندانی بر نتیجه ی کار ما نخواهد داشت چون خود ClustalW میزان این پارامتر را به صورت خودبه خودی تعدیل می کند.
شکاف extension penalties-	این پارامتر اندازه ی شکاف را تعیین می کند. واضح است که تعیین میزان بهینه ی این پارامتر و پارامتر بالایی امکان پذیر است، ولی قاعدتا این میزان برای هر خانواده ی پروتیینی متفاوت خواهد بود و رسیدن به آن نیازمند تجربه می باشد.

جدول ۵-۹: کنترل اثرات تنظیم پارامتر های مختلف در ClustalW

به یاد داشته باشید که هرگز پارامترهای ClustalW را در مورد alignment ی که از درستی آن مطمئن هستید، تغییر ندهید. اگر نتیجه ی کارتان یک alignment ی است که از درستی آن اطمینان دارید به جای تغییر پارامترهای ClustalW، از ویرایشگرهایی که در فصل ۱۰ معرفی شدند برای تصحیح توالیهایتان استفاده کنید. تنها دلیل خوبی که می توان برای تغییر این پارامترها داشت این است که می خواهیم ببینیم با تغییر آن ها آیا نتیجه ی بهتری به دست خواهیم آورد و یا نه؟

چگونه با استفاده از T-coffee توالی ها و ساختارها را هم تراز کنیم؟

T-coffee یک ابزار جدید برای ساخت یک multiple alignment می باشد. هر چند این ابزار شباهت بسیاری با clustalW دارد ولی از جهاتی نیز تفاوتهایی در آن دیده می شود. مثلا این ابزار alignment هایی دقیقتر، با صرف مدت زمان بیشتری را در اختیار ما قرار می دهد.

این ابزار مانند ClustalW، alignment هایی پیشرفته تهیه می کند با این تفاوت که قطعه ها را در تمام طول توالی با هم مقایسه می کند. جدول ۶-۹ کاربردهای اصلی t-coffee را فهرست کرده است. به جز دقت و صحت کار، ویژگی های

دیگری که می توان برای این ابزار برشمرد شامل: توانایی آن در هم تراز کردن توالی ها و ساختارها(EXPRESSO)، امکان ارزیابی صحت یک Alignment (CORE) و امکان ترکیب کردن multiple sequence های گوناگون با یکدیگر و تبدیل آن به یکی (Mcoffee) است.

امکانات	توصیف
T-coffee	ساخت یک multiple sequence alignment.
CORE	میزان اعتبار و صحت یک multiple alignment را ارزیابی می کند.
Mcoffee	همه ی داده های خروجی را در یک alignment نهایی خلاصه می کند.
EXPRESSO	همه ی اطلاعات ساختاری ممکن را در توالی ها جا می دهد. به کمک این امکان می توان بهترین multiple alignment را ساخت.

جدول ۶-۹: ابزار های در دسترس در www.tcoffee.org

چگونه با استفاده از T-coffee یک multiple sequence alignment بسازیم؟

ساخت یک multiple alignment به کمک T-coffee تنها نیاز به کپی کردن توالی ها در پنجره ی مخصوص دارد.

اکنون که توالی های مورد نیاز خود را در دست دارید نوبت به t-coffee می رسد!

۱. وارد سایت www.tcoffee.org شوید.

۲. گزینه ی regular را از خط اول انتخاب کنید تا صفحه ی مربوط به ساخت یک multiple alignment ظاهر شود.

۳. توالی های خود را در پنجره مربوطه paste کنید. در این قسمت می توانید از اکثر فرمت ها استفاده کنید. اگر توالی های خود را به صورت یک فایل text دارید می توانید با انتخاب گزینه ی brewers آن را upload کنید.

۴. بر گزینه ی submit کلیک کنید.

T-coffee بسیار کند می باشد. اگر می خواهید بعد از اتمام محاسبات از نتایج آگاهی یابید، می توانید آدرس پست الکترونیکی خود را در قسمت web form وارد کنید.

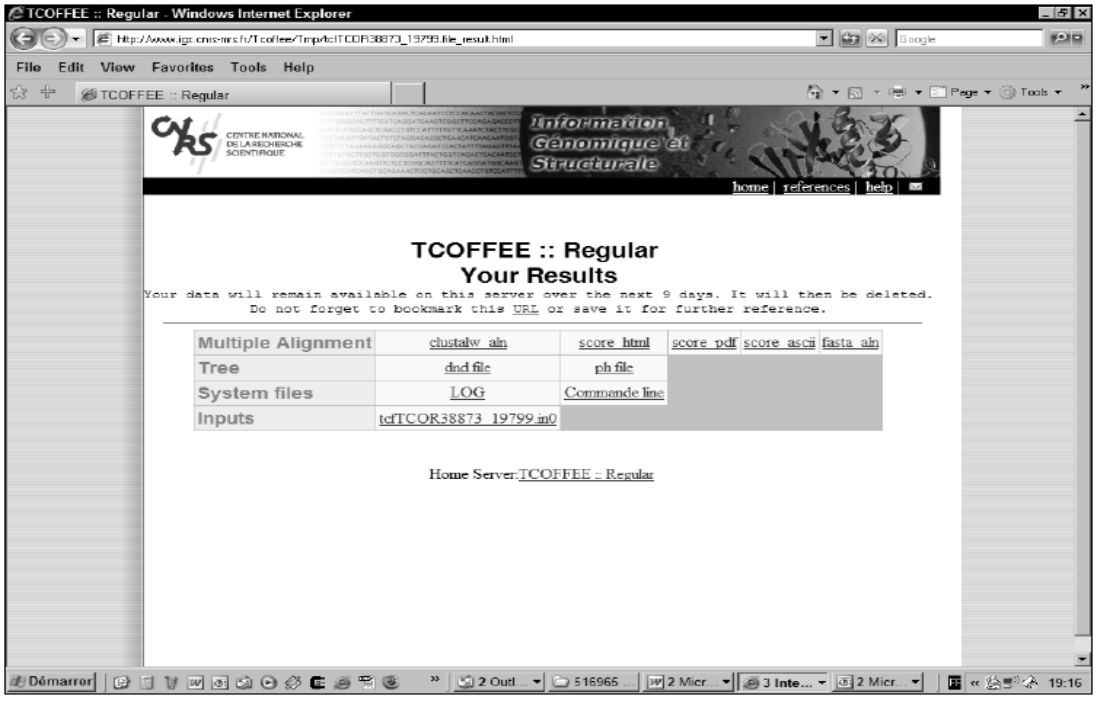
۵. نتایج به دست آمده را بیازمایید.

T-coffee لینک هایی که نتایج شما می باشد را در اختیار شما قرار می دهد (شکل ۸-۹).

ردیف اول جدول مربوط به Multiple sequence alignment می باشد که حاوی موارد زیر است:

- ✓ Msf-aln, clustalW-aln, fasta-aln: فایل هایی متنی که هر کدام مربوط به alignment شما با فرمت های گوناگون می باشد. اگر می خواهید از این داده ها به عنوان داده های ورودی برای برنامه های دیگر استفاده کنید بهتر است که آن ها را نگه دارید.
- ✓ Score-html, score-asii: یک alignment رنگی که هر قسمت بنا بر کیفیت به رنگ خاصی درآمده است. رنگ قرمز مناطقی با کیفیت بالا و رنگ آبی مناطقی با کیفیت پایین را نشان می دهند. Score-html, score-asii فرمت text از فایل html می باشند.

شکل ۷-۹: صفحه Tcofee Build a Multiple Alignment



شکل ۸-۹: خروجی پیش فرض Tcoffee

این دو فایل تنها برای نمایش دادن داده ها مفید هستند و نمی توان از آن ها به عنوان داده ی ورودی برای برنامه های دیگر استفاده کرد.

ردیف دوم جدول مربوط به درخت فیلوژنیک می باشد:

- ✓ dnd: یک درخت راهنما که توسط t-coffee با فرمت newick تهیه می شود (فصل ۱۳ را ببینید). به یاد داشته باشید که هیچ گاه نمی توان از این درخت به عنوان یک درخت فیلوژنیک واقعی استفاده کرد.

- ✓ درخت فیلوژنیک: یک درخت فیلوژنیک واقعی با فرمت newick با استفاده از روش neighbor joining (فصل ۱۳) در t- coffee ساخته می شود.این درخت یک درخت راهنما نیست بلکه یک درخت فیلوژنیک حقیقی است.
- ✓ Pdf: تصویری با فرمت pdf از درخت فیلوژنیک.

چگونه با استفاده از EXPRESSO توالی ها و ساختارها را با یکدیگر ترکیب کنیم؟

EXPRESSO در واقع نوعی t-coffee پیشرفته می باشد که جایگزینی برای 3d-coffee محسوب می گردد.هنگام استفاده از EXPRESSO، این برنامه با استفاده از سیستم BLAST در PDB(پایگاه اطلاعاتی مرتبط با ساختار توالی ها) به دنبال توالی های مرتبط با توالی شما می گردد و سپس از توالی های یافته شده برای ساخت alignment مورد نظر شما استفاده می کند. Alignment ی که بر اساس شباهت ساختارها به وجود آمده باشد مطمئنا دقیق تر و صحیح تر از روش ساده ی ساخت sequence alignment ها هستند.

هرچند که EXPRESSO کندتر از t-coffee عمل می کند ولی اگر ساختارهای مناسب و کافی را بیابد، قابل اطمینان ترین alignment ممکن را ایجاد خواهد کرد. پس ارزش یک بار امتحان کردن را دارد! EXPRESSO برای هم تراز کردن ساختارها از SAP – برنامه ی ساخت Taylor و Orengo- و برای هم تراز کردن توالی ها با ساختارها از FUGUE- بسته ای ساخت Kenji Mizuguchi- استفاده می کند.

برای استفاده از EXPRESSO، به راحتی بر روی گزینه regular – که در خط مربوط به EXPRESSO در سایت t-coffee وجود دارد- کلیک کنید. ادامه ی راه کاملا شبیه به t-coffee است. به قدری کار با این دو برنامه مشابه است که شما به عنوان یک کاربر هرگز متوجه نخواهید شد که اکنون در حال استفاده از ساختارها هستید!

در صفحه ی نتایج فایلی به نام template-list وجود دارد. این فایل حاوی تمام ساختارهایی است که برنامه برای ساخت alignment شما استفاده کرده است. اگر این صفحه خالی بود، یعنی هیچ ساختاری برای هم تراز کردن با توالی های شما یافت نشده و alignment حاضر صرفا یک alignment استاندارد t-coffee می باشد.

چگونه کیفیت alignment خود را با استفاده از CORE ارزیابی کنیم؟

اگر مایل باشید می توانید Multiple sequence alignment خود را که با روش مورد علاقه ی خود ایجاد کرده اید را به t-coffee ارائه کنید و از آن بخواهید که کیفیت آن را برایتان ارزیابی کند. به عبارتی به شما بگوید که کدام قسمت ها معتبر بوده و کدام قسمت ها بهتر است که حذف شوند. برای انجام این کار alignment خود را در سرور CORE- در سایت t-coffee- کپی کنید. برای این کار می توانید از هر کدام از فرمت های رایج (msf, aln, fasta, pir) استفاده کنید.

این گونه ارزیابی یک نوع ارزیابی تجربی است و هرگز نمی تواند جایگزین عدد E شود. با این وجود دانستن اینکه قسمت های با زمینه ی زرد یا قرمز و نارنجی اندکسی بالای ۵(از ۱۰) دارند و بیش از ۸۰٪ مناسب هستند می تواند مفید باشد.

چگونه با استفاده از MUSCLE پایگاه های داده ای بزرگ را کوچک کنیم؟

Muscle را می توان یک تازه وارد در عرصه ی Multiple sequence alignment به حساب آورد. این برنامه می تواند به شما هنگامی می خواهید چند صد توالی را همتراز کنید کمک کند.کار با این برنامه بسیار اسان است. تنها کاری که باید انجام هید این است که توالیهایتان را در آن کپی کنید (www.drive5.com/muscle).

چگونه Multiple sequence alignment خود را تفسیر کنیم؟

تفسیر یک alignment در واقع یک هنر محسوب می شود. عدد E – مقیاسی که میزان قابل اعتماد بودن جستجوی شما را نشان می داد – در زمینه ی Multiple sequence alignment هنوز وجود ندارد. این بدان معناست که برای تفسیر یک Multiple sequence alignment باید از یکسری حدسیات استفاده کرد.

تفسیر کردن alignment های DNA به مراتب سخت تر می باشد. در آنالیز این توالی ها به میزان بالایی از مناطق حفاظت شده نیاز است. ستون های حفاظت شده ی منفرد به احتمال زیاد بی معنی خواهند بود. برای تفسیر این alignment ها به چندین ستون مشابه نیاز است. حتی با وجود توالی های بسیار مشابه هم ساخت یک DNA Alignment کاری دشوار است. به همین دلیل است که بیشتر بیولوژیست ها هنگامی که در انتخاب آزاد هستند، ترجیحا توالی های پروتیینی را انتخاب می کنند.

قسمت های مناسب در یک alignment پروتیینی کدامند؟

راههای مناسبی که ما برای ارزیابی multiple alignment های پروتیینی داریم بیشتر ناشی از اطلاعاتی است که پیرامون ساختار پروتیین ها داریم(بخش ۱۱). ما می دانیم که ساختارهای پروتیینی دارای بخش هایی به نام loop هستند، که این نواحی قسمت های سست ساختار محسوب می شوند که به نواحی محکم که همان نواحی مرکزی هستند متصل می شوند. Loopها به علت سستی به راحتی باز می شوند در حالی که مناطق محکم توالی ها میزان شکاف کمتری دارند. بنابراین مناطق پر شکاف احتمالا مربوط به loop ها بوده و مناطقی با شکاف کم احتمالا مربوط به مناطق مرکزی می باشند.

وقتی به یک alignment خروجی از ClustalW , muscle و یا t-coffee نگاه می کنیم متوجه یکسری علایم به ظاهر رمزی در آن ها می شویم:

- ✓ (*) به یک ستون کاملاً حفاظت شده اشاره دارد.
- ✓ (:) به ستونی اشاره دارد که تمام قسمت های آن اندازه و هیدروپاتی یکسانی دارند.
- ✓ (.) به ستون هایی اشاره دارد که اندازه و هیدروپاتی آن ها در بحث تکامل تغییر نیافته و کاملا حفظ شده است اشاره دارد.

وقتی از این علایم استفاده کنید متوجه می شوید که چقدر ارزشمند هستند. برای مثال نمونه ای از یک توالی مناسب حاوی حدود ۱۰ – ۳۰ آمینواسید، یک تا ۳ ستاره، چندتایی (:) و تعداد زیادی هم (.) می باشد.

موضع جالبی که در زمینهMultiple sequence alignment وجود دارد این است که وجود ۵ -۴ منطقه حفاظت شده در بین حدود ۵۰ آمینواسید به عنوان یک alignment مناسب می تواند کاملا متقاعد کننده باشد. این میزان از ۱۰٪ شباهت هم کمتر است!!! در حالی که قبلا گفتیم که برای در نظر گرفتن دو توالی به عنوان توالی های مشابه به حداقل ۲۵٪ شباهت نیاز داریم.

شما می توانید به راحتی علت قدرتمند بودن Multiple sequence alignment را دریابید. Multiple sequence alignment ها می توانند به شما در یافتن مناطق مشابه در twilling zone ها – مناطقی که شباهت دو به دو هیچ اطلاعاتی در زمینه همولوگوس بودن ذاتی نمی دهند- کمک کنند.

ملاک دیگری برای تعیین یک Multiple alignment به عنوان یک Multiple alignment مناسب نوع اسیدآمینه هایی است که انتظار می رود محافظت شده باشند. اسیدهای آمینه هم اندازه و برابر نیستند و هر کدام از آن ها الگوی مشخصی از جهش و حفاظت در یک Multiple alignment دارند.

یک ستون حفاظت شده تنها زمانی معنی دار است که ستون های کناری آن جهش یافته باشند(محافظت نشده باشند).

Amino Acid	*Characteristic*
W,Y, F	It is common to find conserved tryptophans. Tryptophan is a large hydrophobic residue that sits deep in the core of proteins. It plays an important role in their stability and is therefore difficult to mutate. When tryptophan mutates, it is usually replaced by another aromatic amino acid, such as phenylalanine or tyrosine. Patterns of conserved aromatic amino acids constitute the most common signatures for recognizing protein domains.
G, P	It is common to find conserved columns with a glycine or a proline in a multiple alignment. These two amino acids often coincide with the extremities of well-structured beta strands or alpha helices. (For more on these structures, see Chapter 11.)
C	Cysteines are famous for making C-C (disulphide) bridges. Conserved columns of cysteines are rather common and usually indicate such bridges. Columns of conserved cysteines with a specific distance provide a useful signature for recognizing pro-tein domains and folds.
H, S	Histidine and serine are often involved in catalytic sites, espe-cially those of proteases. Conserved histidine or a conserved serine are good candidates for being part of an active site.
K, R, D, E	These charged amino acids are often involved in ligand binding. Highly conserved columns can also indicate a salt bridge inside the core of the protein.
L	Leucines are rarely very conserved unless they're involved in protein-protein interactions such as a leucine zipper.

جدول ۷-۹: الگوهای محافظت در Multiple Sequence Alignment

روی Multiple Sequence Alignment بیشتر کار کنیم

تمام چیزی که ما از یک Multiple alignment می خواهیم را می توان در یک جمله خلاصه کرد: موقعیت های مهم کدامند؟ درواقع ما از Multiple alignment می خواهیم آمینواسیدهایی که اجازه جهش را نداشته اند، تشخیص دهیم، یعنی مناطقی از توالی ها که در پروتیین های خیلی دور از هم نیز دستنخورده باقی مانده اند را مشخص کنیم. چنین alignment ی را می توان از سرورهای گوناگون مانند t-coffee و یا clustalW به دست آورد.

Alignment ی مناسب هست که شامل پروتیین های مرتبط و در عین حال دور از هم باشد. چنین alignment ی می تواند داستان کاملی از مولفه های گوناگون خانواده ی پروتیینی مورد نظر ما را تعریف کند!

برای مثال با توجه به تحقیقات به اثبات رسیده است که انتهای آمینی نسبت به انتهای کربوکسیلی بیشتر تحت حفاظت بوده است. در انتهای آمینی مناطق کوتاه حفاظت شده زیادی وجود دارد که این ناحیه را به ناحیه مناسبی برای جایگاه فعال بودن بدل کرده است.

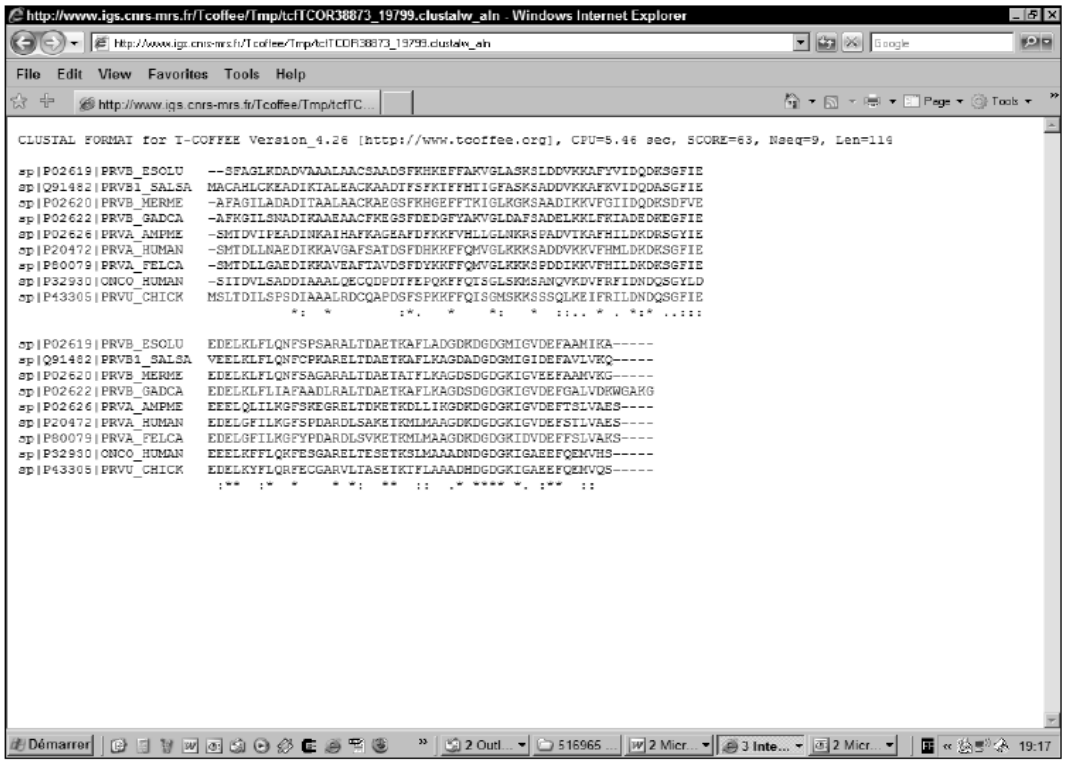
هرچند این نکته جالب است و برای ساخت یک تاریخچه بزرگ کافی نیست! این alignment حاوی مناطق حفاظت شد ی زیادی می باشد که جای پژوهش دارند. کاری که در این زمینه می توان کرد این است که توالی های مرتبط با پروتیین های همین خانواده ولی کمی دور را یکی یکی به مجموعه اضافه کنیم تا در نهایت ببینیم که آیا تغییری در وضعیت alignment ایجاد می کند و یا نه؟ هدف ما از این کار این است که مطمئن شویم که همه ی اعضای خانواده از همین الگو پیروی می کنند.

در این قسمت یک استراتژی احتمالی برای آشکار ساختن عوامل ایجاد کننده تکامل در پروتئین ارائه شده است. اساس این استراتژی ترکیب توالی های همان Multiple Sequence Alignment ی است که BLAST هنگام جستجوی همولوگ های پاروالبومین انسانی در Swiss-Prot، به عنوان marginal hits به شما ارائه کرده است (اگر این مطالب را متوجه نمی شوید، به قسمت "انتخاب توالی ها در سرور ExPASy" در اوایل همین فصل مراجعه نمایید). ابتدا توالی های خود را به صورت زیر جمع آوری کنید:

۱. وارد سایت www.expasy.ch/sprot/sprot-retrievelist.html شوید.
۲. در ردیف format گزینه ی FASTA radio را انتخاب کنید.
۳. شماره ی مسلسل توالی های خود را در پنجره مربوطه وارد کنید.

در هر خط یک شماره را وارد کنید.در مثال ما شماره ها P20472، P80079، P02626، P02619، P43305، P32930، Q91482، P02620، P02622 و P02586 هستند.

P02586 یک TPCS-RABIT تروپونین C خرگوشی است. در BLAST پاروآلبومین انسانی در سرور Swiss-Prot –که ما برای انتخاب سایر توالیهایمان از این سرور استفاده کرده ایم– این توالی با عدد E = ۵ گزارش می شود. این گزارش به نوبه ی خود گزارش چندان خوبی نیست، ولی اکنون که ما در یک Multiple alignment بررسی می کنیم می توانیم این نکته را مورد توجه قرار دهیم که آیا پروتیین خرگوش می تواند اطلاعاتی در مورد پروتیین انسانی در اختیار ما قرار دهد و یا نه؟



شکل ۹-۹: یک multiple sequence alignment خوب

۴. گزینه ی Creat FTP file را کلیک کنید.
۵. توالی های خود را در clipboard کپی کنید.

وقتی تمام توالی ها را آماده کردید اکنون آماده انجام بعضی کارهای مربوط به alignment هستید:

۱. وارد سایت www.ebi.ac.uk/clustalW/index.html شوید.
۲. توالی هایی که در مراحل مقدماتی جمع آوری کرده اید را در پنجره مربوطه کپی کنید.
۳. از منوی alignment گزینه ی fast را انتخاب کنید.
۴. از منوی output format فرمت مورد نظرتان را برای داده های خروجی تعیین کنید.
۵. از منوی output order داده های ورودی را تعیین کنید.
۶. کزینه Run را کلیک کنید.
۷. نتایج به دست آمده را ذخیره کنید.

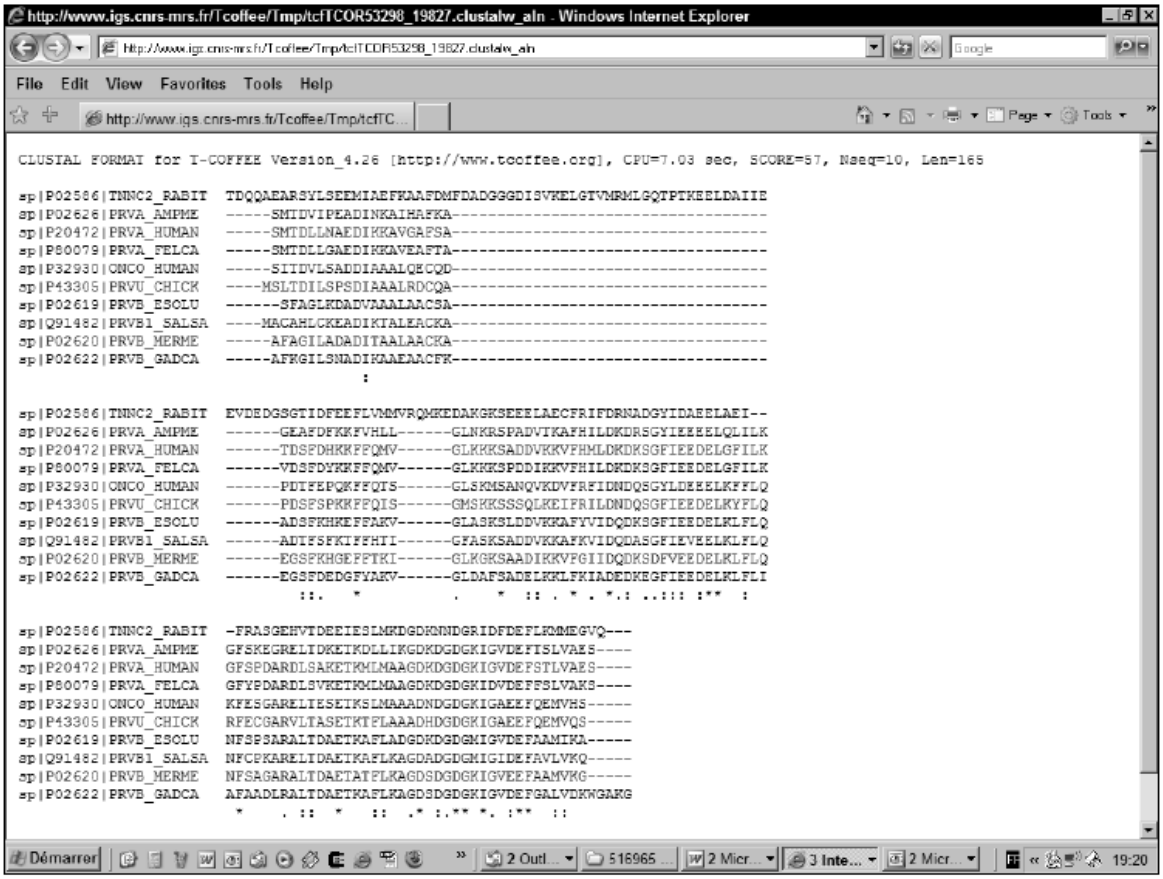
در نتیجه ای که به دست می آید(شکل ۱۰-۹) می توانید ببینید که توالی های جدید بعضی مناطق حفاظت شده را به همان صورت (شکل ۹-۹) دارند در حالی که بعضی از مناطق حفاظت شده را جابه جا کرده اند. نتیجه ی به دست آمده همچنین مناطق مستعد برای حذف و اضافه را نیز به ما نشان می دهد. چنین مناطقی کاندیداهای مناسبی برای مناطق حلقه به حساب می آیند. (مناطق حلقه مناطقی از توالی ها بودند که نسبت به بقیه مناطق توالی سست تر و قابل انعطاف تر بودند. این مناطق در ساختار توالی ها ب هعنوان مناطق اتصال دهنده عمل می کنند. برای یافتن اطلاعات بیشتری در این زمینه به بخش ۱۱ مراجعه کنید.)

شکل ۱۰-۹ همچنین نشان می دهد که نتایج ۲ بخش حفاظت شده را در ناحیه انتهای آمینی نشان می دهند. این نکته می تواند نتیجه ی مهمی به شمار آید. در واقع اگر به دنبال اسیدآمینه های مسئول عملکرد پروتیین هستیم می توانیم با توجه به این یافته تا حدی این احتمال را درست تلقی کنیم.

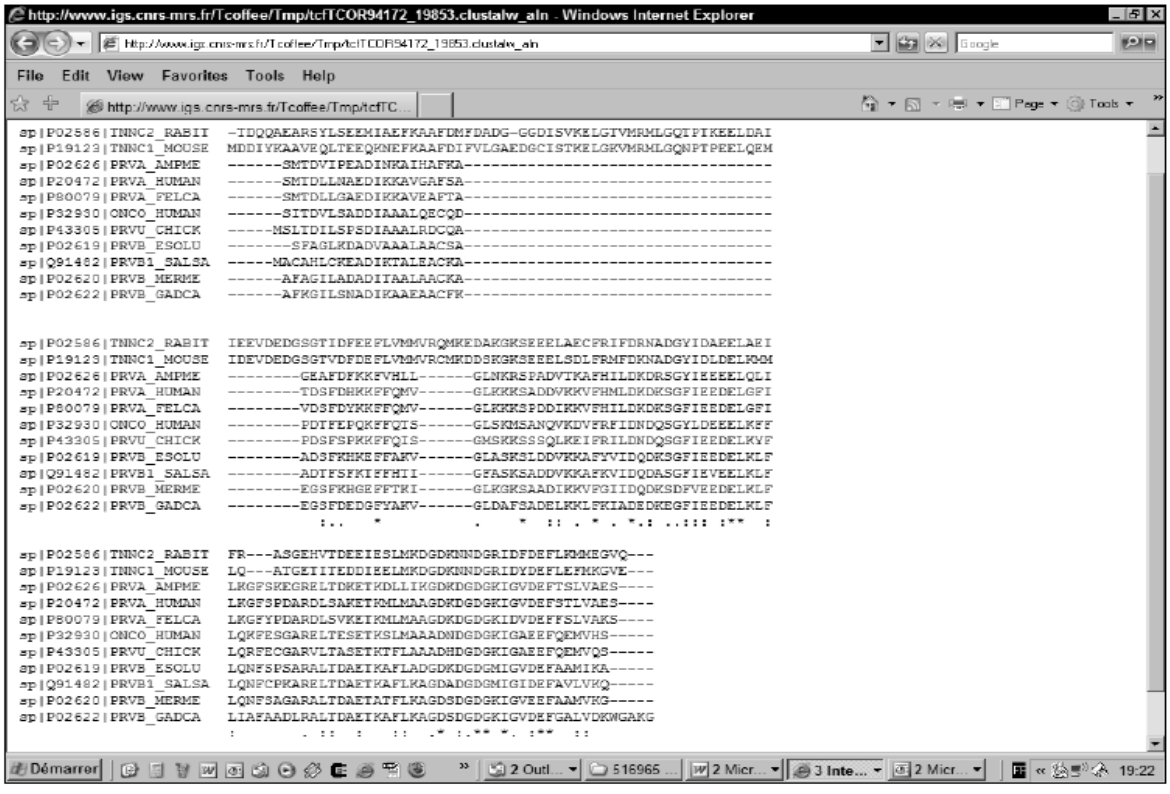
برای به اثبات رساندن این فرضیه می توانیم این روند را از ابتدا تکرار کنیم منتها با این تفاوت که یک توالی دیگر از همین خانواده منتها اندکی دورتر را نیز به توالی های قبلی اضافه کنیم.

در اینجا ما p19123 یک TPCC_MOUSE که تروپونین C موش می باشد را اضافه می کنیم.این توالی به مقدار کاملا جزیی با پاروو‌آلبومین انسانی همولوگ می باشد. شکل ۱۱-۹ نتیجه وارد شدن این پروتئین جدید به multiple sequence alignment ما را نشان می دهد. این شکل به وضوح نشان می دهد که اکثریت مناطق محافظت شده باقی مانده اند.

با چنین نتیجه ای می توانیم با اطمینان بگوییم که این دو منطقه ی محافظت شده در بعضی عملکردهای بیولوژیکی درگیر هستند. با توجه به این که بیشتر این پروتیین ها به کلسیم متصل می شوند می توانیم چنین نتیجه بگیریم که جایگاه اتصال کلسیم احتمالا بعضی از این مناطق حفاظت شده را در بر دارد.این نتیجه با نتیجه ای که از Swiss-Prot به دست آورده بودیم مطابقت دارد.



شکل ۱۰-۹: انجام multiple alignment پروتئین های با ارتباط کم



شکل ۱۱-۹: وارد کردن خویشاوندهای دور تر به multiple alignment

چگونه توالی های که قابل هم تراز کردن نیستند را مقایسه کنیم؟

گاهی پیش می آید که شما در کارتان باید توالی هایی که از یک نیای مشترک نیستند و یا بقدری از هم دور هستند که نمی توان آن ها را همولوگ به حساب آورد را با یکدیگر مقایسه کنید.

برنامه های مرتبط با Multiple sequence alignment در چنین شرایطی بی فایده خواهند بود. همه این برنامه ها DNA و یا پروتیین هایی را شامل می شوند که هرچند ممکن است به ظاهر هیچ ارتباطی با هم نداشته باشند، ولی دارای قلمروهای همولوگ هستند.

در این شرایط انتظار هیچ گونه معجزه ای را از بیوانفورماتیک نداشته باشید. در چنین شرایطی احتمالا به این نتیجه می رسید که هیچ راهی برای انجام این کار وجود ندارد، ولی ما به شما پیشنهاد می کنیم که قبل از انصراف کامل، این ۲ تکنیکی که در زیر به شما آموزش داده می شود را امتحان کنید:

- ✓ Gibbs sampler: به طور همزمان به تمام توالی های شما به صورت مختصر رسیدگی می کند.
- ✓ Pratt: الگوهای انعطاف پذیر را جستجو می کند. دسته ای ویژه از قطعات که می توانند شکاف دار باشند و به حفاظت در موقعیت های خاص و مشخص نیاز داشته باشند.

چگونه با استفاده از Gibbs sampler یک Multiple sequence alignment بسازیم؟

در واقع این برنامه کاملا بر پایه قانون احتمالات است. یعنی برنامه ابتدا توالی ها را بالا و پایین می کند و آن ها را به طورکاملا شانسی با یکدیگر هم تراز می کند تا جایی که به یک راه حل مناسب دست یابد. بعد از آن همین روشی که تا حدودی درست بوده است را ادامه می دهد تا کار خود را بهتر کند. نکته ی جالبی که در این برنامه وجود دارد این است که برای حل مسایل پیچیده لازم نیست که همه ی راه های احتمالی را دنبال کنیم تا به نتیجه برسیم!

Gurs چنین روشی را stochastic نامید. چون در حقیقت عنصر اصلی که در آن نقش دارد عنصر شانس و احتمالات است. این نوع روش ها، روش های بسیار قدرتمندی در علم بیوانفورماتیک به حساب می آیند هرچند که شهرتی که شایستشان می باشد را هنوز به دست نیاورده اند. علت این بی توجهی را شاید بتوان در این نکته خلاصه کرد که اگر ۲ بار از این برنامه در دو مجموعه توالی استفاده کنیم این امکان زیاد است که به نتایج مشابهی دست نیابید! برای بیشتر بیولوژیست ها قابلیت تولید مجدد یک نتیجه بیشتر دارای اهمیت است، و از این نکته متنفر هستند که هر لحظه کامپیوتر نظرشان را تغییر دهد.

با وجود این اگر کمی منطقی فکر کنیم متوجه می شویم که این برنامه می تواند راه حل های مفیدی را برای مسایل بسیار پیچیده ی ما ارائه دهد.برای مثال Gibbs sampler در یافتن محدوده ی HTH(helix-turn-helix) در طول خانواده ی پروتیینی بسیار کارآمد می باشد. همچنین یک راه مناسب برای جستجوی عوامل تنظیمی موجود در توالی های DNA که به هم ارتباطی ندارند نیز می باشد.این برنامه را می توانید در آدرس bayesweb.wadsworth.org/gibbs/gibbs.html و یا در سایت موسسه پاستور bioweb.pasteur.fr/seqanal/interfaces/gibbs-simple.html بیابید.

به یاد داشته باشید هنگامی که از Gibbs sampler استفاده می کنید باید صحیح رفتار کنید. یعنی باید از حداکثر تعداد ممکن توالی برای کارتان استفاده کنید. هرگز از کمتر از ۲۰ توالی استفاده نکنید.

فصل دهم

چگونه **alignment** ها را متناسب با نیازمان تنظیم کرده و برای انتشار عمومی آماده کنیم؟

در این فصل می خوانیم:

- ✓ چگونه می توانیم فرمت توالی های مورد نظرمان را تغییر دهیم؟
- ✓ چگونه به صورت دستی و با استفاده Jalview امکان دستکاری ، تنظیم و تصحیح توالیهایمان را خواهیم یافت؟
- ✓ چگونه alignment ها را به فرمت " لوگو" تبدیل کنیم؟
- ✓ چگونه توالی های همسان را با استفاده از Boxshade تزئین کنیم؟

.....

" ظاهرا، این تفکر که «هر چیز بالفعل عین حقیقت است» تفکر درستی است"

در این بخش چگونگی استفاده و آماده کردن Alignment ها را جهت عرضه به مجامع علمی به شما نشان خواهیم داد. مهم ترین وسیله برای خواندن این فصل، داشتن یک alignment به عنوان نمونه، برای تمرین نکات گفته شده می باشد. به همین دلیل در ابتدای بخش یک alignment ساختگی را برای آن دسته از خوانندگانی که نمونه ای برای تمرین دراختیار ندارند معرفی خواهیم کرد.

در اولین مبحث این فصل، بهتر دیدیم که اطلاعاتی مختصر و مفید پیرامون انواع فرمت ها -که در هر زمان ممکن است با آن ها مواجه شوید- در اختیار شما بگذاریم و در کنار معرفی فرمت ها چگونگی انتخاب فرمتی متناسب با نیازتان را به شما آموزش خواهیم داد و یا به زبان ساده تر، به شما یاد خواهیم داد که چگونه از یک جنگل پر از انواع فرمت ها به سلامتی عبور کنید!!!

اگر با استفاده از روش های اتوماتیک" توالی های هم تراز" را به دست آورده باشید، احتمالا قبل از هرگونه استفاده از آن ها، به اِعمال یکسری تغییرات نیاز خواهید داشت. تنظیم و تصحیح توالی ها کار چندان آسان و ساده ای نیست: در جاهایی باید GAP

اضافه کرد، در جای دیگر باید چند توالی را یکباره و به طور کامل جابه جا کرد و یا ....وقتی چنین تغییراتی را با یک واژه پرداز معمولی انجام دهید، بعد از گذشت چند ساعت چنان دچار سردرگمی خواهید شد که به مثابه درگیری با یک پازل ۱۰۰ تایی خواهد بود. این که با چه آرامش و دقتی کار را شروع کرده باشید چندان مهم نیست، بالاخره بعد از چند دقیقه همه چیز از کنترلتان خارج خواهد شد. برای جلوگیری از چنین اغتشاشاتی، در این فصل سرورهای اینترنتی مناسبی جهت تصحیح و تنظیم توالیهایتان به شما معرفی خواهد شد تا دیگر مجبور به استفاده از واژه پرداز های معمولی و یا برنامه هایی مشابه آن ها نشوید.

احتمالا هدف شما بعد از یافتن یگ گروه رؤیایی از توالی های متشابه، گزارش آن به جوامع علمی بزرگ است. هیچ چیز در نظر دیگران زیباتر و بهتر از یک شمای زیبا ، با انتخاب رنگ مناسب به همراه چند صفحه تفسیر و توضیح مختصر پیرامون آن – که وجود چنین تفسیرهایی بسیار هم مهم است- نیست. در واقع Alignment هایی با ظاهر مناسب به قدری متقاعد کننده هستند که حتی خود شما با دیدن آن ها بسیار راحت تر از زمانی که هیچ تزئینی نداشتند از کیفیت بالای آنها مطمئن خواهید شد . ما در این بخش سرورهای اینترنتی به شما معرفی خواهیم کرد که چنین خدماتی را به راحتی برای شما انجام دهند.

در پایان بخش، ۲ ابزار مناسب را به شما معرفی خواهیم کرد که با استفاده از آن ها به راحتی می توانید اطلاعات مورد نظرتان را از Multiple alignment ها به دست آورید. در واقع مزیت قابل تأمل این دو پایگاه این است که از Protein logos استفاده می کنند.

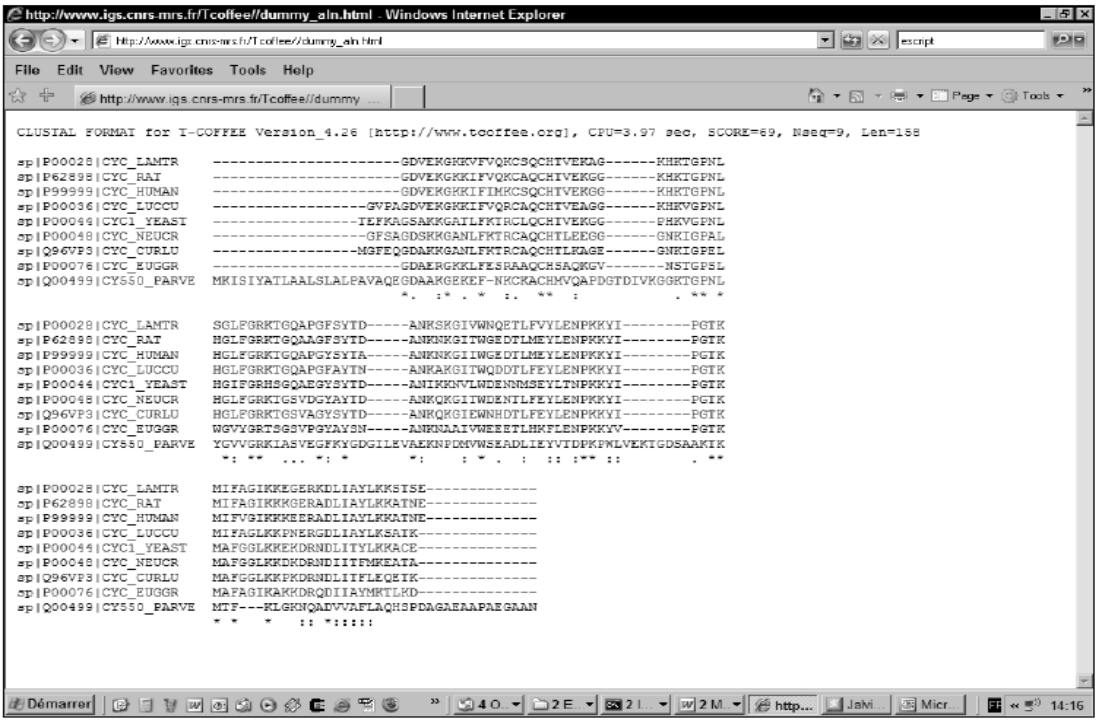
Logo فرمتی از نمایش اطلاعات است که داده ها را بر اساس میزان حفاظت شدگی اسیدهای آمینه نشان می دهد(شکل۱۵). این فرمت به تفصیل در انتهای بخش توضیح داده خواهد شد.

به طور کلی، فصل پیش رو، فصلی کوتاه است ولی کوتاهی آن دال بر مختصر بودن نیست. ما در این بخش تمام اطلاعات لازم برای ساخت فرمتی مناسب و قابل فهم جهت گزارشات عمومی را به شما نشان خواهیم داد. بنابراین دیگر هیچ بهانه ای برای گزارش Alignment هایی ساده، خام و غیر قابل فهم نخواهید داشت!

اگر اکنون Alignment ی آماده برای کار در این فصل ندارید وارد سایت زیر شوید:
http://www.tcoffee.org/dummy_aln.html
و از alignment موجود در آن با فرمت ALN استفاده کنید .(شکل شماره ۱ – ۱۰).

بهترین فرمت، متناسب با نیاز ما، کدام است؟

داده های ورودی و خروجی از یک برنامه ی آنلاین – و حتی هر برنامه ی دیگر- به طور ۱۰۰٪ در محدوده ی کنترل ما نیست. برای مثال داده های خروجی از اکثر برنامه های مرتبط با توالی های متشابه به صورت فرمت MSF(Multiple sequence format) می باشند.حال در نظر بگیرید که فرمت داده های شما به صورت MSF است، ولی برنامه ای که برای تجزیه و تحلیل آن ها در اختیار دارید تنها فرمت FASTA را میشناسد. چه میکنید؟ آیا راهی وجود ندارد و باید کار را از ابتدا به صورتی دیگر تکرار کنید؟ نه چنین نیست. بهترین و راحت ترین کار استفاده از یک برنامه ی تغییر دهنده ی فرمت است که در این بخش نمونه ای از آن به شما معرفی خواهد شد.



شکل ۱-۱۰: یک multiple sequence alignment در فرمت ALN یا ClustalW

هنگام کار با برنامه های Sequence alignment با گزینه ای تحت عنوان " بهترین فرمت" و یا چیزی مشابه با آن مواجه نخواهید شد. کسی که با برنامه های اینترنتی کار کرده باشد حتما به این نکته پی برده است که هر سرویس اینترنتی، با فرمتی خاص کار می کند و این انتخابی نیست که در اختیار ما گذاشته شده باشد.(برای توضیحات بیشتر کادر " فرمتی برای همه: دموکراسی و یا آنارشیسم؟" را مطالعه کنید). به دست آوردن اطلاعاتی جامع پیرامون انواع فرمت ها چیزی است که تنها با تکرار و تمرین به دست خواهد آمد. ولی ما در این بخش به شما چندین ایده در این زمینه خواهیم داد و در ضمن آن، چشم اندازی از مشکلات موجود را به برای شما ترسیم خواهیم کرد.

یک فرمت عمومی برای همه: دموکراسی و یا آنارشیسم

در علم بیوانفورماتیک، تنوع در فرمت ها یکی از مشکلات بسیار بزرگ محسوب می شود. در گذشته سنت بر این پایه استوار بود که طراحان برنامه ها، فرمتی تازه و مختص خود ایجاد کنند. هر فرمت جدید اندکی متفاوت با فرمت های موجود و متناسب تر با گروه هدف خود بود. برای مثال دانشمندان فیلوژنی فرمت Phylip را ترجیح می دهند و یا استفاده کنندگان GCG(بسته ای اطلاعاتی مرتبط با علم بیوانفورماتیک) فرمت MSF را بیشتر مورد استفاده قرار می دهند و بقیه رشته ها نیز به همین صورت. چه ما وجود انواع فرمت ها را دوست داشته باشیم و چه نه، فرقی نخواهد کرد. این قانون تاریخی باعث به وجود آمدن انواعی از فرمت های قابل قبول شده است که امروزه همه آن ها در کنار یکدیگر وجود دارند و مورد استفاده قرار می گیرند.

عامل دیگری را که می توان به عنوان منشأ این پیچیدگی به شمار آورد، کشمکش میان بیولوژیست ها و دانشمندان کامپیوتر است. طی سالیانی طولانی بیولوژیست ها با زبان کامپیوتر مشکل داشتند و حرفشان این بود که:" ما این زبان سنگین کامپیوتر را نمی توانیم درک کنیم". بنابراین در مقابله با مهندسین کامپیوتر – و البته بنا بر نیازشان – فرمت هایی متناسب با کار خود ایجاد کردند. درست است که دانشمندان هر دو رشته تا حدودی محق هستند، ولی این حق به حل این مشکلی کمکی نکرد.

با آمدن فرمت XML همه چیز تغییر کرد. XML(Markup Language eXtensible) فرمتی مشابه با HTML(زبان اینترنت) است. در واقع این زبان، داده های شما را به فرمتی قابل قبول برای همه ترجمه می کند. امروزه دانشمندان بیولوژیست و کامپیوتر بر این نکته اتفاق نظر دارند که این زبان می تواند راه حلی برای مشکل مذکور باشد به طوریکه امروزه برنامه های بیوانفورماتیک به صورتی طراحی شده اند که اکثرا داده های خروجی خود را بر اساس آن تنظیم می کنند.

متأسفانه هنوز هم استاندارد مشخصی تعیین نگردیده است و تعداد سرورهایی که از فرمت XML استفاده می کنند بسیار ناچیز است . در واقع دور از انتظار نیست که در دهه ۹۰ زندگی نوه هایمان نیز هنوز از این فرمت های قدیمی استفاده کنند!!!

قبل از اینکه فرمتی خاص را برای بایگانی داده های خود انتخاب کنید بهتر است ابتدا این ۴ سوال را از خود بپرسید:

- ✓ آیا اکثر برنامه های موجود این فرمت را می شناسند؟
- ✓ آیا تمام همکاران من نحوه ی استفاده از این فرمت را می دانند؟
- ✓ آیا تمام داده های مورد نیاز خود را می توانم با این فرمت ذخیره کنم؟
- ✓ آیا داده هایی با این فرمت قابلیت دستکاری را دارند؟

امروزه فرمتی چنین ایده آل وجود ندارد که جوابی مثبت برای تمام سوالات مذکور را در پی داشته باشد. در واقع انتخاب ما باید متکی بر " نیازهای واقعیمان" و " امکانات موجود" باشد. تا جایی که ممکن است نباید از چندین فرمت به صورت مخلوط استفاده کنیم.

طبقه بندی انواع فرمت های مرتبط با sequence alignment نسبتا پیچیده است. لیستی از تمام فرمت های محتمل را در جدول ۱- ۱۰ می توانید ببینید.

فرمت موجود	نوع فرمت	مورد استفاده ی آن
Post- script, PDF, HTML	گرافیکی	فرمتی نهایی که تنها برای انتشار مناسب است.
FASTA	متنی	دستکاری در آن به راحتی امکان پذیر است. قابل خواندن توسط انسان نیست. تغییر نام توالی ها آسان است. برای بسیاری از برنامه ها قابل فهم است. امکان اضافه کردن در آن وجود ندارد.
PIR	متنی	بسیار مشابه با FASTA است و تنها تفاوتی که با آن دارد وجود تعدادی خط اضافی برای اضافه کردن مقداری محدود اطلاعات به آن

		است.
MSF	متنی	فرمتی بسیار استاندارد برای Multiple- alignment. اعمال تغییر و دستکاری در آن سخت است. برای انسان قابل فهم است. امکان اضافه کردن در آن وجود دارد. برای بسیاری از برنامه های موجود تعریف شده، است. (البته بعضی از برنامه ها آن را به صورت محدود و ناحیه ای پشتیبانی می کنند!)
Selex	متنی	فرمتی عمومی تر از MSF است. امکان اضافه کردن اطلاعات بیشتر به آن وجود دارد. برای تعداد کمی از برنامه ها تعریف شده است.
ALN	متنی	مدلی ساده از MSF است. فرمت پیش فرض برای ClustalX است برای انسان قابل فهم است. برای بسیاری از برنامه های موجود تعریف شده است.
Phylip	متنی	نوع دیگری از فرمت ALN است. برای تجزیه و تحلیل های فیلوژنیک مناسب است. برای بسیاری از برنامه های فیلوژنیک تعریف شده است.

جدول ۱-۱۰: دسته بندی فرمت های مختلف multiple sequence alignment

هر چند فرمت گرافیکی فرمت چندان مناسبی جهت تجزیه و تحلیل نیست ولی فرمتی شایع در انتشار Alignment هاست. اگر به یاد داشته باشید منابع ارئه دهنده این فرمت از alignment ها را در بخش قبلی همین فصل به شما نشان دادیم. در ابتدای این فصل فرمت ALN (جدول ۱-۱۰) به شما معرفی شد. فرمت FASTA نیز بارها در فصول قبلی نام برده شده است. درکارهای آزمایشی بیشترین فرمتی که مورد استعمال قرار می گیرد فرمت MSF می باشد. به این نکته توجه داشته باشید که هر چند فرمت ALN و MSF فرمتی مناسب برای استفاده انسان به نظر می آید ولی خواندن آن برای ماشین ها کار چندان آسانی نیست.

فرمت های اصلی موجود کدامند؟

با وجود اینکه فرمت غالب در نمایش Alignment ها، فرمت گرافیکی است ولی فرمتی مناسب و ارزشمند برای تجزیه و تحلیل محسوب نمی شود. در آخرین بخش از این فصل، منابعی به شما معرفی خواهد شد که Alignment هایی با فرمت گرافیکی در اختیارتان قرار می دهند. در این بخش همچنین فرمت های ALN(شکل ۱- ۱۰) و FASTA (فرمتی که بارها در بخش

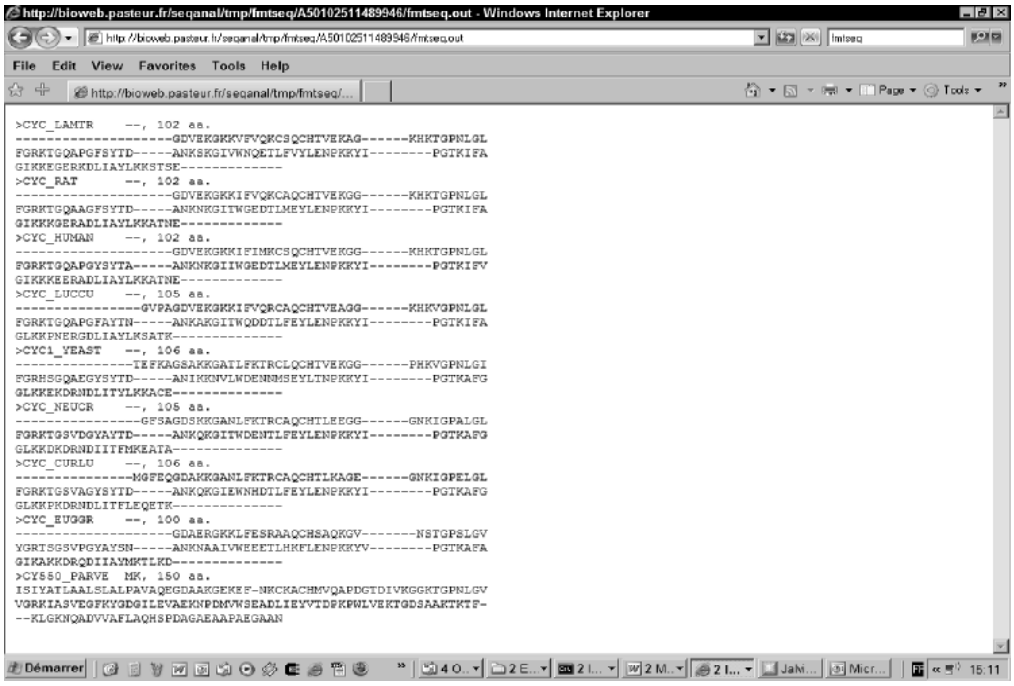
های قبل به آن اشاره شد. شکل ۳ – ۱۰) به شما نشان داده خواهد شد. فرمت غالبی که معمولا در کارهای تحقیقاتی مورد استفاده قرار می گیرد فرمت MSH است(شکل ۲ – ۱۰). بار دیگر بر این نکته تأکید می کنیم که همیشه به یاد داشته باشید: هر چند فرمت های ALN و MSH فرمت های قابل فهمی برای انسان به حساب می آیند ولی درکشان برای ماشین ها کار آسانی نیست در حالی که فرمت FASTA دقیقا عکس آن هاست.

به یاد داشته باشید که برای ایجاد یک فرمت قابل فهم و خواندن برای انسان بهترین روش تقسیم کردن توالی های همتراز چندگانه به بلوک های ۶۰ تایی است. بدین ترتیب بلوک ها یکی پس از دیگری به نمایش در می آیند.

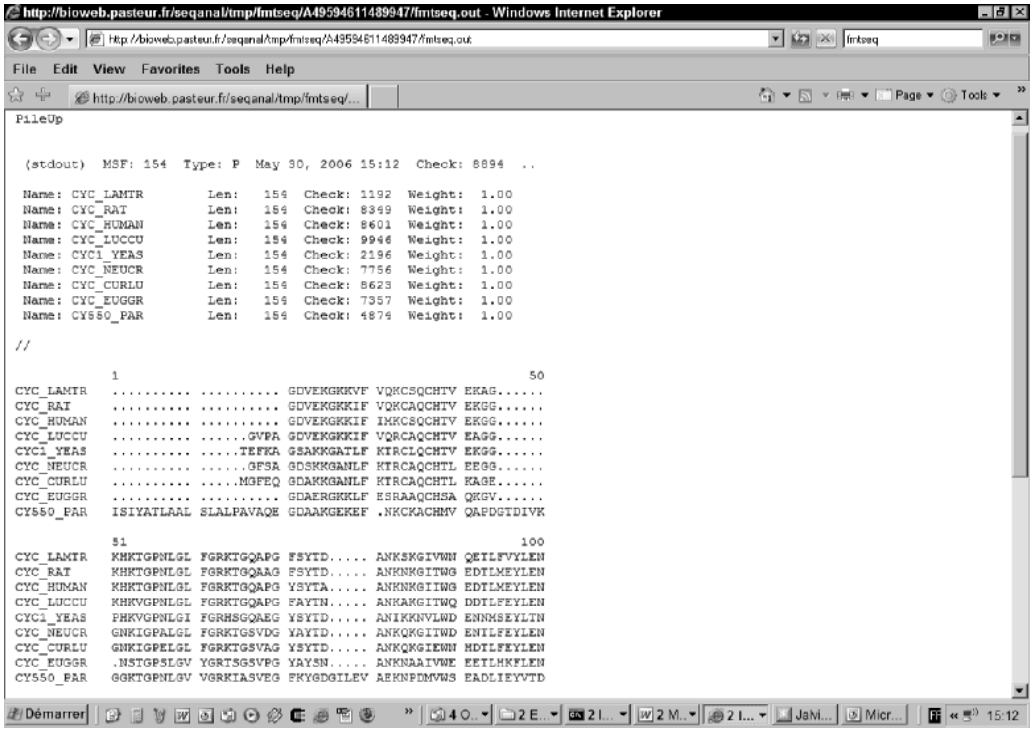
اگر اطلاعات بیشتری در مورد دنیای فرمت ها نیاز دارید می توانید از دو سایت زیر کمک بگیرید: http://apps.sourceforge.net/trac/sitedocs/wiki و http://www.ebi.ac.uk/help/formats.html

کدام فرمت بهترین انتخاب ماست؟

گاهی یک سرور داده های خروجی خود را با انواعی از فرمت ها در اختیار ما قرار می دهد. در این زمان شاید فرمت گرافیکی با آن رنگ های زیبا و چشمک زن به نظر ما کامل ترین فرمت بیاید. چنین داده های گرافیکی معمولا با فرمت های post- script, PDF,.html,.gif,.jpeg هستند. اگر فریب رنگ های زیبا و جذاب داده های گرافیکی را بخوریم و تنها آن ها را نگه داشته و بقیه را از بین ببریم، خود را در مرداب بدی گیر انداخته ایم.



شکل ۲-۱۰: multiple sequence alignment در فرمت FASTA



شکل ۳-۱۰: یک multiple sequence alignment در فرمت MSF

داد های گرافیکی مدلی بسیار مناسب برای نمایش اطلاعات موجود هستند هرچند هنوز تجزیه و تحلیل یک داده ی گرافیکی در برنامه های مخصوص تجزیه و تحلیل داده ها، غیر ممکن است. نگهداری یک شمای گرافیکی به عنوان یک نقاشی زیبا معقول به نظر می رسد ولی به شما توصیه می کنیم که علاوه بر آن حتما یکی از انواع فرمت های متنی مربوط به نتایج کارتان را هم نگه دارید. تبدیل فرمت متنی به فرمت گرافیکی بسیار آسان تر از تبدیل معکوس آن است.

گزارش نتایج کار به صورت فرمت متنی بسیار بهتر و علمی تر است. در واقع فرمت متنی فرمت پیش فرض در بسیاری از برنامه هاست. هدف اصلی از انتخاب یک فرمت، انتخاب فرمتی است که قابل نمایش باشد، به راحتی قابل دست کاری و تصحیح باشد، و برای سرورهای دیگر هم تعریف شده باشد تا اگر زمانی لازم به استفاده از سرور دیگری بود، نیاز به تکرار تمام مراحل قبلی در سرور جدید نباشد. البته واضح است که فرمت های متنی بسیار ساده تر از انواع گرافیکی خود هستند!

MSF فرمتی بسیار استاندار و مناسب برای (sequence alignment Multiple) است. به عبارتی در هنگام دانلود یک Alignment از اینترنت انتخاب فرمت MSF کار عاقلانه ایست. FASTA و یا PIR نسبت به MSF عمومیت کمتری دارند ولی آن ها نیز نسبتا مناسب هستند. در هنگام استفاده از برنامه های فیلوژنیک، فرمت Phylip مناسب تر است. پیش فرض خروجی در سرور ClustalW – گسترده ترین برنامه برای Multiple sequence alignment- ALN است. بنابراین این فرمت را نیز می توان به عنوان یکی از فرمت های پرکاربرد و مشهور به حساب آورد. در واقع ALN می تواند فرمتی مناسب برای جایگزین شدن بسیاری از فرمت های موجود باشد.

چگونه یک فرمت را به فرمتی دیگر تبدیل کنیم؟

اگر برنامه ی مورد استفاده ی شما فرمتی متناسب با نیازتان را در اختیارتان قرار نداد بهترین کار استفاده از یک برنامه ی " تبدیل کننده" است تا فرمت موجود را به فرمت مورد نظر شما تبدیل کند. در این قسمت شما یاد خواهید گرفت که چگونه فرمت ALN (شکل۱-۱۰) را به فرمت FASTA تبدیل کنید.

نام برنامه	آدرس اینترنتی برنامه	توضیحات پیرامون برنامه
Fmtseq	http://bioweb.pasteur.fr/seqanal/interfaces/fmtseq-simple.html http://www.bimcore.emory.edu/pise/	توانایی تغییر بسیاری از فرمت ها را دارد.
READSEQ	http://www-bimas.cit.nih.gov/molbio/readseq/ http://dot.imgen.bcm.tmc.edu/seq-util/readseq.html http://iubio.bio.indiana.edu/cgi-HYPERLINK "http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi"bin/readseq.cgi http://www.ebi.ac.uk/cgi-bin/readseq.cgi	یک برنامه ی بسیار مشهور و عمومی در تغییر دادن فرمت است ولی متأسفانه فرمت ALNرا نمی شناسد.
SeqCheck	http://darwin.nmsu.edu/bioinfo/seqcheck/seqcheck.php	توالی هایی با فرمتFASTA را پاک می کند.

جدول ۲-۱۰: تغییر فرمت در اینترنت

مراقب مفقود شدن داده هایتان باشید!!!

در پروسه ی تغییر فرمت همیشه امکان از دست دادن مقداری از اطلاعات وجود دارد. هر فرمتی، نوع خاصی از اطلاعات را نگه خواهد داشت. مشکل اصلی این است که در بسیاری از مواقع وقتی از مفقود شدن اطلاعاتمان مطلع می شویم، که برای انجام هر کاری دیر شده است. جدول ۳ - ۱۰ لیستی از تمام اطلاعاتی که در این پروسه امکان مفقود شدنشان وجود دارد را در اختیار شما قرار می دهد.

البته تمام سرورها در حذف داده ها مشابه هم عمل نمی کنند(حتی سرورهایی که عنوان مشترکی دارند مثلا READSEQ). ۲ سرور از READSEQ می توانند انواع متفاوتی از مشکلاتی که در جدول زیر ذکر شده است را ایجاد کنند. بنابراین می توان چنین بیان کرد که مشکلاتی که یک سرور برای داده های ما ایجاد می کند با مشکلات ایجاد شده توسط یک سرور دیگر می تواند به طور کامل متفاوت باشد. تمام این توضیحات تنها برای این بود که همیشه چشمانتان را باز نگه دارید و یک کپی از تمام داده های اصلی خود را ذخیره داشته باشید تا مجبور به تکرار تمام مراحل کاریتان نشوید!!!.

داده هایی که امکان از دست رفتنشان وجود دارد	ماهیت و دلیل بروز خطا
نام توالی	در هنگام تبدیل یک فرمت به فرمتی دیگر، در فرمت های که نام بلندی دارند امکان بروز اشکالی در نامشان وجود دارد .بالاخص امکان تغییر حروف وجود دارد .چنین اتفاقی معمولا در هنگام تبدیل فرمت FASTA به فرمت ALN بروز می کند. نتیجه ی چنین تغییراتی غیر قابل پیش بینی است .گاهی اوقات حتی امکان دارد تکه ای از نام توالی به خود توالی اضافه شود.
توالی با کیفیت بالا/ پایین	FASTAتنها شامل اطلاعاتی با کیفیت بالاست. بیشتر برنامه ها از دریافت اطلاعاتی متفاوت از نظر کیفیت، نفرت دارند!
نحوه ی نمایش GAPها	بعضی از انواع فرمت ها مانندMSF ، برای نشان دادنGAP از علایمی متفاوت نسبت به سایر فرمت ها استفاده می کنند .)، -، .(~معمولا در پروسه ی تغییر فرمت، این علایم به صورت شکل روبرو(_) در خواهند آمد.
تفاسیر و اطلاعات اضافه بر برنامه	MSFتوانایی تعیین عدد وزنی توالی ها را داردکه چنین داده هایی در اکثر برنامه های تبدیل کننده از بین خواهد رفت. هنگام تغییر فرمت از PIR معمولا اطلاعات اضافی موجود در آن از بین خواهد رفت. در فرمت FASTA اطلاعاتی که بعد از نام توالی می آیند، به طور حتم با تغییر فرمت از بین خواهند رفت.
آمینو اسید و یا نوکلئوتید خاصی	بعضی از برنامه های تغییر فرمت برای موارد مبهم، کدی را جایگزین می کنند. برای مثال به جای آمینواسیدهای نامشخص" X "و برای نوکلئوتیدهای نامشخص" N "را جایگزین می کنند. اگر ما پی در پی توالی های خود را از یک برنامه به برنامه ی دیگر بفرستیم ممکن است چنین کدهایی از بین بروند و این زمانی که ما بر روی تمام اسیدهای آمینه موجود حساب کرده ایم مشکل زا خواهد بود. اگر این امکان را دارید تنها از ۴ نوع نوکلئوتید و ۲۰ نوع اسید آمینه ی استاندارد استفاده کنید.

جدول ۳-۱۰: اطلاعاتی که هنگام تغییر فرمت امکان دارد از دست بروند

Jalview وسیله ای برای تنظیم Multiple sequence alignment

طراحی متدهای مرتبط با Multiple sequence alignment بدون نقص نبوده است و سازندگان این برنامه ها خود به تمامی این نقوص واقفند، با این وجود استفاده کنندگان نیز باید این کمبود ها را شناسایی کرده تا در هنگام استفاده دقت و توجه کافی مبذول دارند.

برای ایجاد یک Alignment مطابق با چیزی که مورد نظرمان است باید ساعت ها خود را با یک برنامه درگیر کنیم تا شاید نتیجه ای مشابه با چیزی که می خواهیم به دست آوریم. چنین کاری شاید ساعت ها زمان ببرد وهرگز هم نتیجه ی ای مطابق با خواسته مان را در پی نداشته باشد. حتی ممکن است تصمیم بگیریم نتایج به دست آمده از یک برنامه را به صورت دستی تغییر دهیم و برای این کار از برنامه هایی مانند Word Microsoft استفاده کنیم. چنین تجربه ای ممکن است حتی افراد بسیار باهوش و زیرک را نیز سردرگم کرده و در نهایت بدون اینکه به نتیجه ای رسیده باشیم داده های ابتدایی خود را نیز از دست بدهیم. پس توصیه می کنیم هرگز چنین کاری را انجام ندهید.

تصحیح و تغییر توالی های چندگانه کار چندان آسانی نیست. بنابراین بیولوژیست ها در پی آن برآمده اند که برنامه هایی تحت عنوان "Text editing" ایجاد کنند. با استفاده از این برنامه ها ما به راحتی و بدون سردرگمی تغییرات لازم را اعمال می کنیم. چنین برنامه هایی به راحتی تمام مراحلی که در حال انجام بر روی توالی های شما می باشد را نشان می دهند و ضمن آن، این امکان را برای شما فراهم می کنند که به راحتی توالی هایتان را به زیر مجموعه هایی تبدیل کنید. در نتیجه، دیگر اجباری به تغییر یک توالی به طور کامل نخواهید داشت و تنها می توانید بر روی دو زیر مجموعه کار کرده و آن ها را با هم تطابق دهید. اکثریت این برنامه ها قبل از اولین استفاده، خواستار نصب نرم افزارهایی بر روی کامپیوترتان هستند. بنابراین وقتی با چنین چیزی روبرو شدید از ادامه ی کار نهراسید و دستورات لازم را انجام دهید. بسیاری از این برنامه ها – مانند Seaview_ به راحتی قابل نصب بر روی کامپیوترهای خانگی هستند.

باز هم اگر نمی خواهید چنین کاری را انجام دهید، می توانید از Jalview استفاده کنید. برای استفاده ی فوری از چنین برنامه هایی Jalview مناسب است، چون تنها با نصب آن بر روی سرور امکان استفاده را پیدا خواهید کرد.

Jalview مشابه با Dotlet (فصل ۸) و با همان ویژگی هاست:

- ✓ بر روی کامپیوتر شخصی شما قابلیت اجرا خواهد داشت.
- ✓ هنگامی که توالی مورد نظرتان را بر روی Jalview دانلود می کنید این توالی بر روی کامپیوتر شما خواهد ماند. این بدان معنی است که کس دیگری به توالی شما دسترسی نخواهد داشت و شما با امنیت می توانید کار خود را انجام دهید.

اگر می خواهید از لو نرفتن اطلاعاتتان در اینترنت مطمئن شوید، به محض بارگزاری Jalview از منوی File گزینه Work Offline را انتخاب کنید.

قبل از انجام دادن این مراحل هرگز اطلاعات محرمانه ی خود را بر Jalview بارگزاری نکنید. ارتباطات اینترنتی هرگز امنیت 100%را تضمین نمی کنند.

شروع کار با Jalview

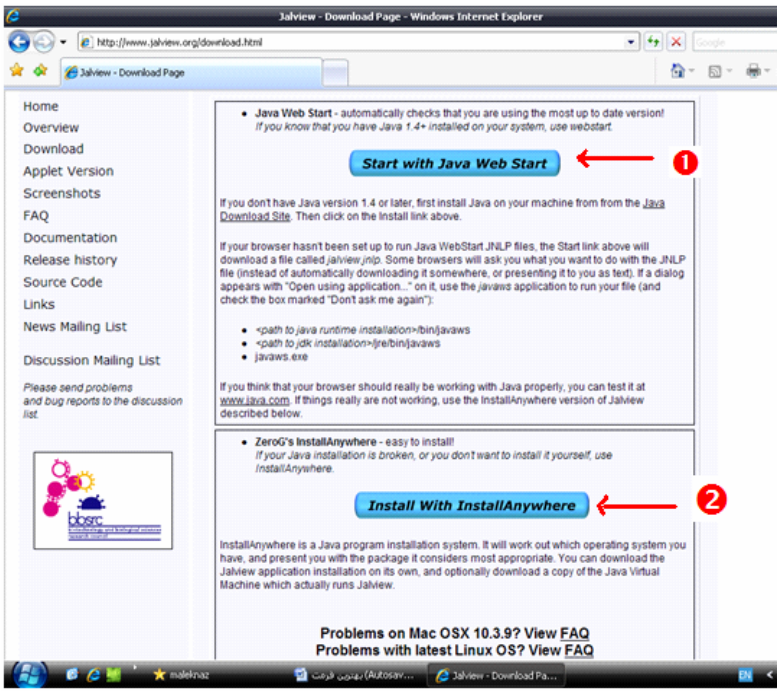
ما در این بخش تنها اطلاعاتی پایه ای و مقدماتی در مورد امکانات و موارد استفاده از Jalview را در اختیار شما قرار خواهیم داد. اگر به اطلاعات بیشتری در این زمینه نیاز داشتید می توانید از قسمت Help این سرور استفاده کنید.

۱. آدرس زیر را وارد کنید:

http://www.jalview.org/download.html

۲. برنامه ای به نام - JAVA که احتمالا نام آن برایتان آشنا می باشد- برای کار با Jalview لازم است. و علاوه بر آن یک نصب کوتاه از خود برنامه ی Jalview نیز مورد نیازست.

در صفحه ای که بعد از وارد کردن آدرس بالا باز شد، برای فراهم آوردن این اطلاعات ۲ گزینه ی پیشنهادی را در برابر شما قرار می دهد. در گزینه ی اول(شکل ۴-۱۰) پیش فرض آن است که شما برنامه ی java را بر روی کامپیوترتان آماده دارید و تنها به نصب برنامه مربوط به خود jalvew نیاز دارید. بنابراین اگر چنین است، گزینه ی مربوطه را انتخاب کنید تا مقدمات ابتدایی کارتان به راحتی فراهم شود. ولی اگر Java را هم بر روی کامپیوتر شخصیتان ندارید، نگران نباشید. با انتخاب گزینه ی دیگر(۴-۱۰ شماره ۲) برنامه ی Java و jalview توام با یکدیگر نصب خواهد شد. در طی نصب، هر گزینه ای که نیاز به موافقت شما دارد را انتخاب کنید.



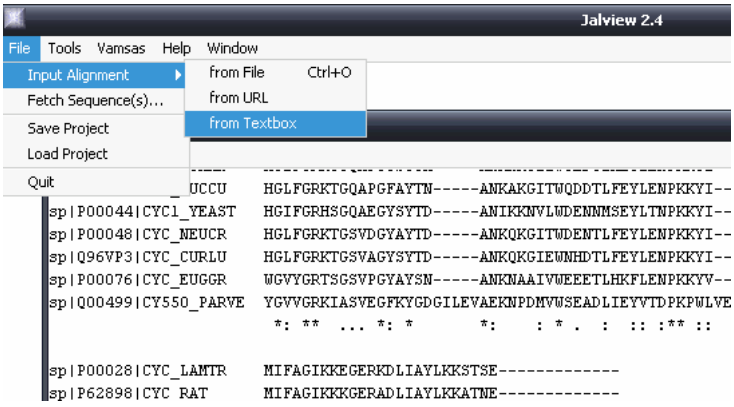
شکل ۱۰ – ۴

اکنون تمام وسایل مورد نیاز برای شروع کارتان آماده هست. پس با خیال راحت ادامه مراحل را طی کنید:

بعد از نصب این برنامه ها بر روی کامپیوترتان، دیگر نیازی به نصب مجدد در دفعات بعدی نمی باشد. برای استفاده در دفعات بعدی تنها کافی است بر روی همان گزینه ی ۱ در شکل ۵ کلیک کنید. به این صورت ابتدا وجود چنین برنامه ای بر روی کامپیوتر شما چک خواهد شد و بعد از چند ثانیه صفحه ی اصلی، جلوی چشمانتان خواهد بود.

۳. مطابق شکل ۶ گزینه ی File input alignment from textbox را انتخاب کنید.

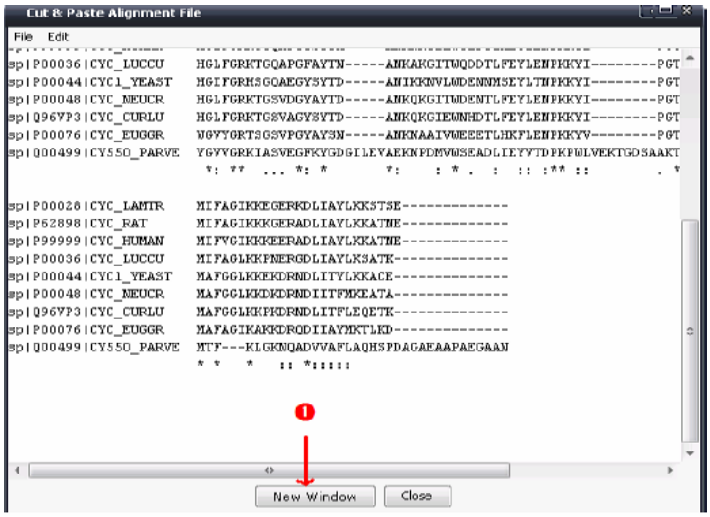
اگر alignment مورد نظرتان را در یک فایل به صورت آماده دارید، می توانید با انتخاب گزینه ی from file به جای from textbox از آن استفاده کنید.



شکل ۵-۱۰

۴. بعد از انجام مراحل گفته شده در مرحله ی پیشین، صفحه ای سفید با نام Cut And Paste Alignment File باز خواهد شد. Alignment مورد نظر خود را از منبع اصلی کپی و در این صفحه ی سفید paste کنید.(شکل۵-۱۰)

اگر alignment خاصی مورد نظرتان نمی باشد از همان Alignment ی که در آدرس http://www.tcoffee.org/dummy_aln.html وجود دارد استفاده کنید.



شکل ۶-۱۰

با کلیک بر گزینه ی new window (شکل۶-۱۰) صفحه ای جدید با عنوان "input cut and paste" باز می شود که در آن alignment مورد نظرتان کاملا آماده برای تنظیم شدن به دست شماست!!!!(شکل۷-۱۰).

۵. گزینه ی ClastalX را از منوی color انتخاب کنید(شکل۷-۱۰).

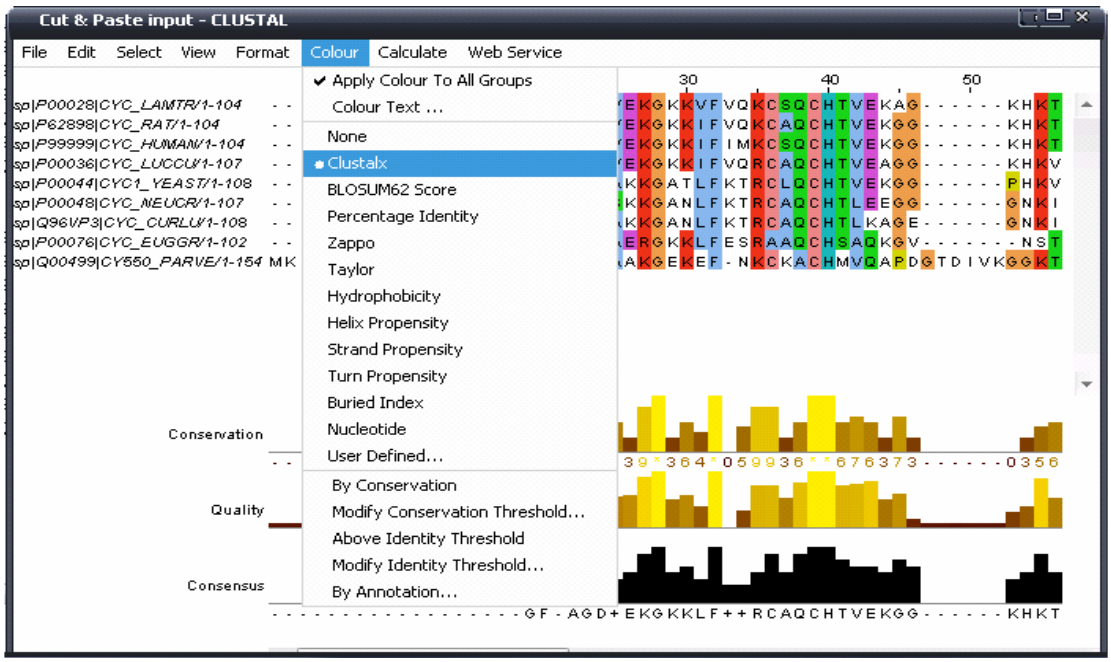
شما می توانید بر اساس سلیقه ی شخصیتان یکی از انواع طرح های موجود برای نشان دادن alignment را انتخاب کنید. هر فرد آشنا به این برنامه، می تواند ساعت ها برای علت انتخاب یکی از انواع این طرح ها برایتان سخنرانی کند. پس می بینید که انتخاب هرکدام از این گزینه ها می تواند مفهومی در پی داشته باشد!!!

در هنگام نشان دادن alignment تان به دیگران، باید بتوانید به این سوال عادی بیولوژیست ها پاسخ دهید و آن اینست که "طرح مورد استفاده در این Alignment چه بوده است؟"

برای یافتن اطلاعاتی بیشتر در مورد انواع طرح های موجود می توانید از قسمت Help سرور Jalview استفاده کنید.

لازم به ذکر است که تمام انواع طرح های قابل استفاده در این سرور را می توانید در منوی color بیابید.

ClastalX که در این مثال مورد انتخاب ما بود یکی از طرح های رایج به حساب می آید.



شکل ۷-۱۰

چگونه می توانیم به راحتی، دسته ای از توالی ها را به صورت دسته جمعی تنظیم کنیم؟

در تنظیم یک alignment معمولا هدف اصلی، ایجاد تطابق میان چندین توالی است که این توالی ها به هر دلیلی با هم شباهت هایی دارند. تنظیم چند توالی به صورت همزمان را می توان "اصلاح دسته جمعی توالی ها" نامید. قبل از آن که بخواهید چندین توالی را به صورت همزمان تصحیح کنید، باید آن ها را تحت نام یک گروه در بیاورید. مراحل زیر در این زمینه به شما کمک خواهند کرد:

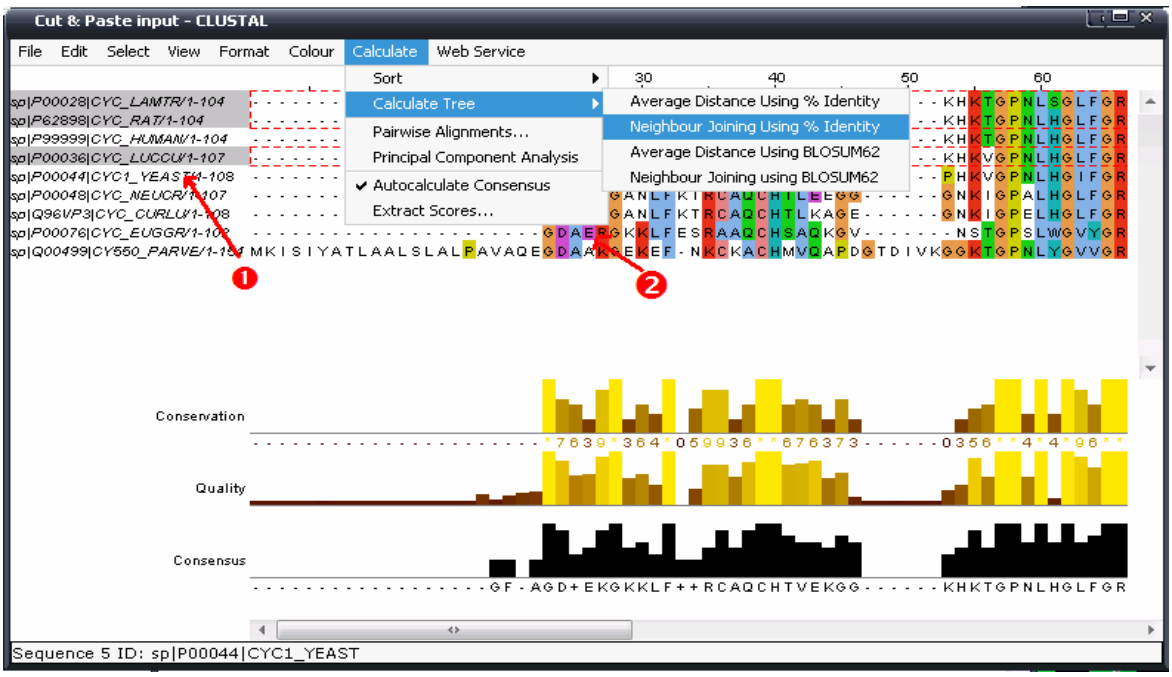
۱. برای دستیابی به هدفی که در بالا به آن ذکر شد، ابتدا مراحل موجود در قسمت قبلی را تکرار کنید.

۲. همزمان با فشار کلید Ctrl توالی ۱ و ۲ و ۴ را انتخاب کنید.(شکل۷-۱۰)

انتخاب چندین توالی با فرمت فیلوژنیک نیز امکان پذیر است:

الف- برای این کار گزینه ی jointing tree using PID treenaighbor calcute calcute را انتخاب کنید(شکل۷-۱۰).

ب- در درخت فرعی که باز می شود می توانید توالی های مورد نظر را به صورت منفرد انتخاب کنید(شکل ۸-۱۰)



شکل ۸-۱۰

برای حذف و یا وارد کردن "GAP" در میان توالی ها:

الف- کلید Ctrl را فشار دهید.

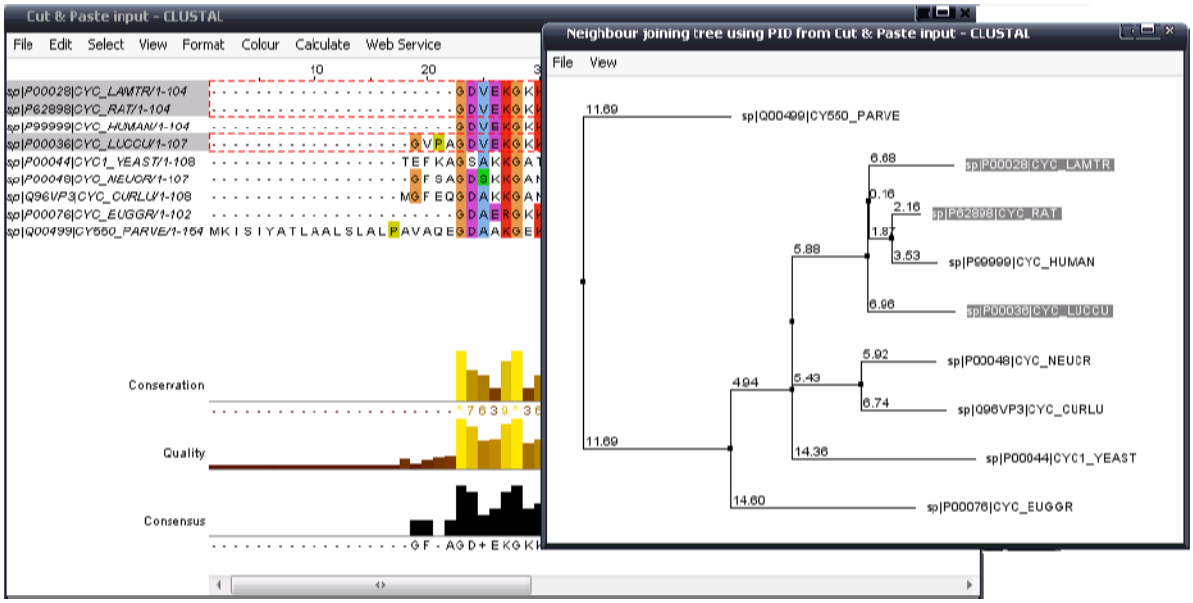
ب- نشانه ی موس خود را در سمت راست محل مورد نظر برای "وارد کردن و یا حذف فاصله" قرار دهید.

ج- موس خود را به سمت راست و یا چپ بکشانید تا به نتیجه ی مطلوبتان دست یابید.

برای حذف و یا اضافه کردن GAP بر روی یک توالی به تنهایی، می توانید به جای استفاده از گزینه ی Ctrl از گزینه ی shift استفاده کنید.

برای حذف ردیف های خالی موجود، از منوی Edit گزینه ی column empty Remove را انتخاب کنید(شکل ۱۰-۱۰، شماره ۱) چنین ردیف های خالی ای می تواند در نتیجه حذف و اضافه کردن Gap ها به وجود آید. با انجام این مرحله آن ها به صورت اتوماتیک برای شما حذف خواهند شد.

با انتخاب Editpad Gap از حذف تمامی ردیف های خالی اطمینان حاصل کنید.(شکل ۱۰- ۱۰، شماره ۲)

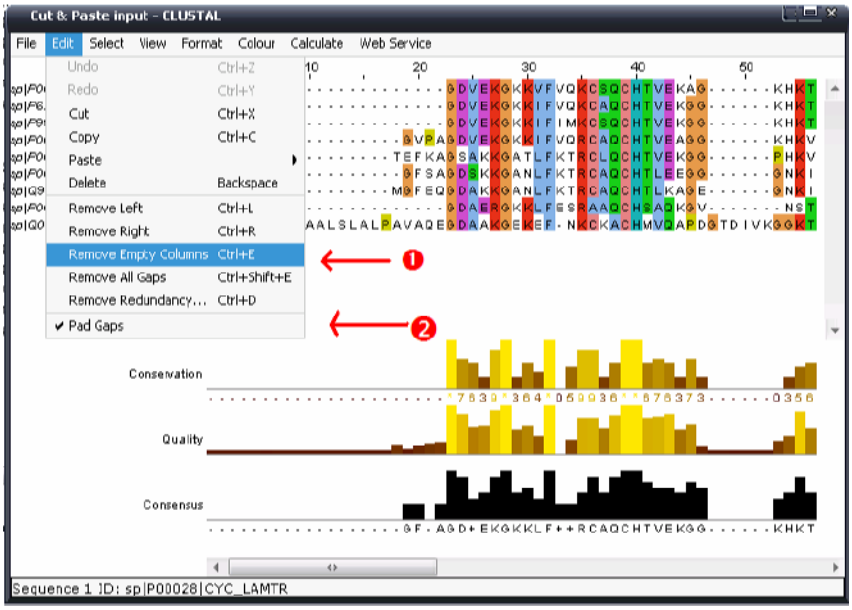


شکل ۹-۱۰

با ایجاد تغییر در مکان توالیها در واقع تعدادی Gap در بین آن ها ایجاد می کنید که این GAPها می توانند در کار شما اختلال ایجاد کنند. مثلا بر اثر وجود آن ها ممکن است توالی های شما هم تراز به نظر نیایند.

برای تنظیم یک alignment دیگر به صفحه ی اصلی Jalview برگردید و تمام مراحل را از ابتدا بر روی آن اجرا کنید.

عملکرد گزینه ی file در منوی اصلی با منوی alignment متفاوت است و در نتیجه ی استفاده از هر گزینه در آن ها با نتیجه ی متفاوتی روبرو خواهید شد.



شکل ۱۰ - ۱۰

ویژگی های مثبت Jalview

هرچند یکی از اهداف اصلی این فصل معرفی کامل امکانات jalview است ولی به نظر ما بهتر است که خود شما هم در مورد امکانات و استفاده های آن جستجو کنید. Jalview در واقع مانند همه ی برنامه هایی است که تا به حال به شما معرفی کرده ایم: قرار نیست با کار کردن با آن، شما چیزی را خراب کنید، پس با خیال راحت تجربه کنید!!!

نابود کردن نتایج به دست آمده از کار بر روی یک alignment کار چندان سختی نیست. پس به یاد داشته باشید که همیشه هر مرحله ای از کارتان را که به اتمام رساندید، ذخیره کنید و هرگز به گزینه ی Undo و یا Ctrl + Z امید نداشته باشید.

جدول ۴- ۱۰ لیستی از موارد مفید را در اختیارتان قرار می دهد.

فرمان که دانستن آن مفید است	این فرمان چه خواهد کرد؟
Calculate Auto calculate Consensus	به صورت خودکار، گرافی های مرتبط با alignmentتان را به روز می کند. در چنین گرافی مرزی مانند شفق میزان تفاوت ها را نشان می دهد. با انتخاب این گزینه، در طول تنظیم توالی ها، گرافتان به صورت اتوماتیک به روز می شود.
Edit Remove Redundancy	اگر درصد خاصی از شباهت، مورد نظرتان است، با انتخاب این گزینه و تعیین میزان مورد نظرتان به نتیجه ی مطلوب خواهید رسید.
Calculate Tree Neighbor Joining Tree Using PID	توالی های موجود را برحسب رابطه ی فیلوژنیک(نژادی) و به صورت یک درخت نمایش می دهد که می توان با انتخاب هر کدام از آن ها گروهی از توالی ها را برای تنظیم انتخاب کرد.

جدول ۴-۱۰: چند ویژگی مفید Jalview

ذخیره داده ها در Jalview

در نسخه جدید jalview ذخیره هر دو نوع داده ی گرافیکی و متنی آسان است. ما به شما پیشنهاد می کنیم که هر دو نوع داده ی خود را ذخیره کنید.

داده های رنگی برای توضیح و نشان دادن اطلاعات مناسبند در حالی که داده های متنی فرمتی مناسب برای ادامه ی کار می باشند.

۱. از گزینه ی فایل در صفحه اصلی Cut and paste input ،گزینه ی save as را انتخاب کنید.
۲. در صفحه ای که باز می شود در قسمت file as type فرمت FASTA را انتخاب کنید.

پیشنهاد ما به شما این است که همیشه یا فرمت FASTA و یا فرمت CLUSTAL (ALN) را برای ذخیره انتخاب کنید. چون خواندن این دو فرمت برای ما انسان ها راحت تر می باشد.

۳. از منوی File گزینه Export و سپس HTML را انتخاب کنید.

با انجام مراحل بالا Alignment شما با فرمتی رنگی ذخیره خواهد شد. این فرمت را با استفاده از انواع Web browser می توانید باز کنید.

چگونه **Multiple alignment** خود را برای انتشار و گزارش عمومی آماده کنیم؟

بالاخره زمان نشان دادن نتایج کارتان رسید: اکنون شما دارای یک alignment هستید که نه تنها خواستار، بلکه مصمم برای نشان دادن ان به جهانیانید. شما می خواهید آن را به همکارانتان نشان دهید، یا شاید هم تصمیم دارید که آن را منتشر کنید و یا حتی بر روی لیوان ها، پوسترها، تی شرت ها و یا دیگر کالای تبلیغاتی به چاپ برسانید. به طور خلاصه شما به یک شکل کوچک اما به اندازه ی کافی گویا برای متقاعد کردن دیگران به شگفت انگیز بودن نتیجه ی کارتان نیاز دارید. در این بخش ما به شما تعدادی از ابزارهای موجود بر اینترنت که می تواند کمکیار شما در حصول نتیجه باشد را نشان می دهیم

Boxshade چیست؟

Boxshade برنامه ی سودمندی است که با اضافه کردن افکت هایی به alignment ها به آن ها جان می دهد. این برنامه alignment ها را بر اساس میزان شباهتی که از نظر محتویات خود(اسیدهای آمینه و یا ژنوم ها) دارند، رنگ آمیزی خواهد کرد و علاوه بر آن فایل هایی را ایجاد می کند که با استفاده از آن ها می توانید توالی های خود را با فرمتی مناسب برای الحاق به مقالات آماده کنید.(مثال هایی دیگر از چنین برنامه هایی در جدول ۷- ۱۰ آمده است)

برای استفاده از این سرور توالی زیر را دنبال کنید:

۱. در نوار جستجو، آدرس زیر را وارد کنید:

http://www.ch.embnet.org/software/BOX_form.html

صفحه ای مطابق با شکل۱۱-۱۰ باز می شود.

۲. از منوی Output format گزینه ی RTF_new را انتخاب کنید.(شکل ۱۱-۱۰، شماره ۱)

اکثریت فرمت های موجود در این منو، قابل استفاده بر روی یک کامپیوتر خانگی نیستند.بهترین فرمت قابل استفاده در این گزینه، فرمت RTF_new می باشد که در نتیجه ی انتخاب آن فایلی با قابلیت خوانده شدن توسط بسیاری از برنامه های واژه پرداز(مانند Microsoft World) ایجاد خواهد شد.

۳. در قسمت Font Size فونت مورد نظر خود را وارد کنید.(۱۱-۱۰ ، شماره ۲)

اگر alignment انتخابی شما طویل می باشد بهتر است فونتی کوچک انتخاب کنید.

اگر در منوی اول فرمت RTF_new را برای داده های خروجیتان انتخاب کرده اید، بعدا هم امکان تغییر فونت آن را خواهید داشت.

۴. از منوی consensus line گزینه ی Add a Consensus Line with Letters را انتخاب کنید.(10-11، شماره ۳)

با انتخاب این گزینه، یک توالی، شامل آمینواسید غالب در هر ردیف، به آن اضافه خواهد شد.

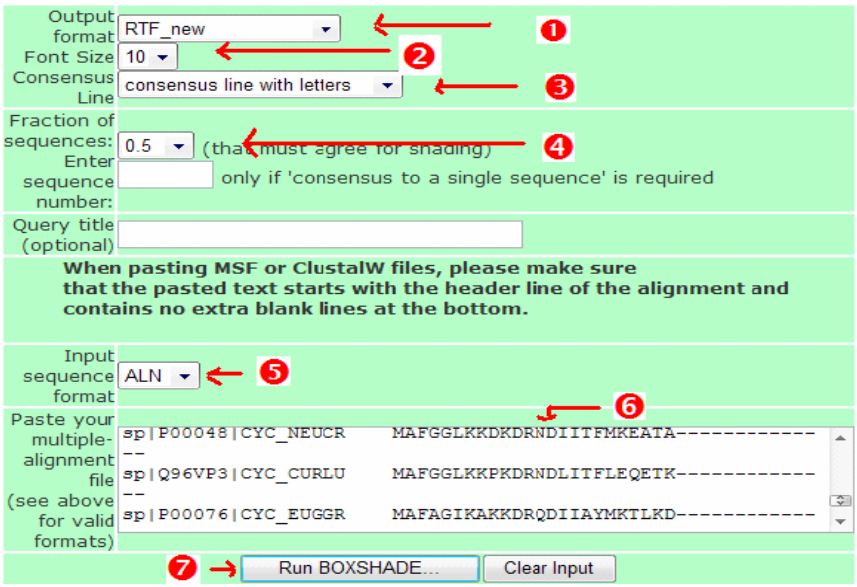
اگر تنها توالی هایی با مقدار مشخصی از شباهت مد نظر شما هستند، با تعیین عدد مورد نظرتان در این بخش به راحتی به هدفتان خواهید رسید (10-11، شماره ۴).

۵. با انتخاب یک عدد، Boxshade تنها ستون هایی را تحت افکت خود قرار می دهد که این میزان از Conservation را دارا باشند.برای مثال اگر در این بخش عدد ۵/. را انتخاب کنید، سیستم تنها توالی هایی را در نظر می گیرد که ۵۰ ٪ از آن ها حاوی یک اسید آمینه و یا یک نوکلئوتید است.

Conservation تنها شامل اسیدآمینه ها و نوکلئوزید های قابل شناسایی برای boxshad نیست. مهم شباهت دو ذره می باشد. مثلا ایزولوسین و والین که نسبت به یکدیگر مشابهند نیز به عنوان conservation محسوب می شوند.

۲ نوع افکت در Boxshad وجود دارد:

- ✓ سیاه: آمینواسیدها و یا نوکلئوتید های کاملا یکسان.
- ✓ خاکستری: آمینواسیدهای مشابه.



شکل 10-11

۶. در نوار input sequence format فرمت alignment خود را معین کنید. (شکل 10-11، شماره ۵) فرمت alignment ما در این جا ALN است.

فرمت های مناسب دیگر برای این جا فرمت هایی است که قابل شناسایی برای برنامه های READSEQ باشد. فرمت هایی مانند FASTA و PIR.

۷. توالی خود را در مکان مشخص شده Paste کنید.(شکل 11-10، شماره ۶)

۸. بر روی گزینه ی run کلیک کنید.(شکل 11-10، شماره ۷)

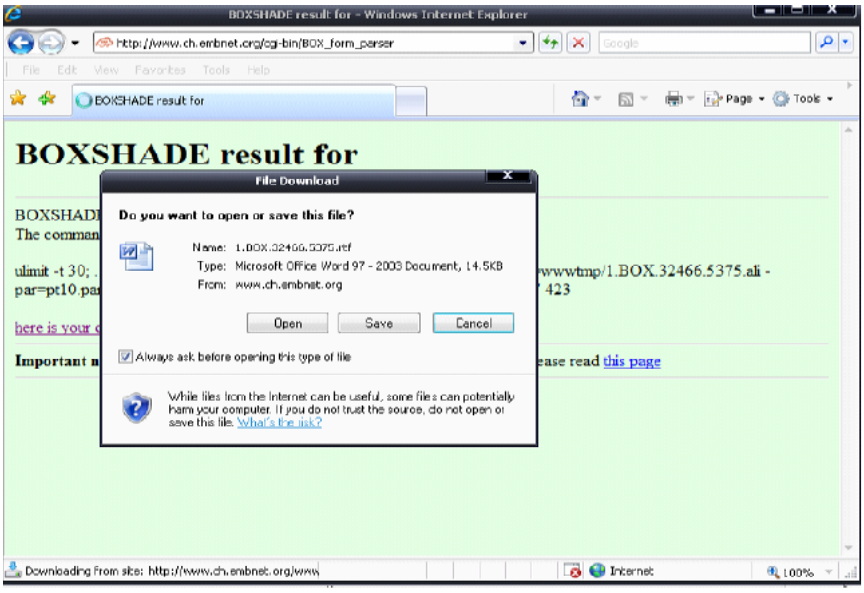
صفحه ای شبیه شکل ۱۲-۱۰ باز می شود.

اگر در این صفحه ، هیچ مورد هایپر لینکی وجود نداشت در واقع بدان معنیست که boxshad در خواندن alignment شما دچار مشکل شده است. در مواقعی که با چنین مشکلی مواجه شدید، بار دیگر فرمت alignment خود را بررسی کنید. شاید اشکال کار از آن جا باشد و حتی اگر نتوانستید مشکلی بیابید، برای اطمینان فرمتی دیگر را انتخاب کنید تا شاید این مشکل حل شود.

۹. بر روی here is your output کلیک کنید. صفحه ی مشابه شکل ۱۳-۱۰ باز می شود.

۱۰. با کلیک بر روی open ، alignment شما در صفحه ی Word باز خواهد شد(شکل ۱۴-۱۰).

شکل ۱۲-۱۰



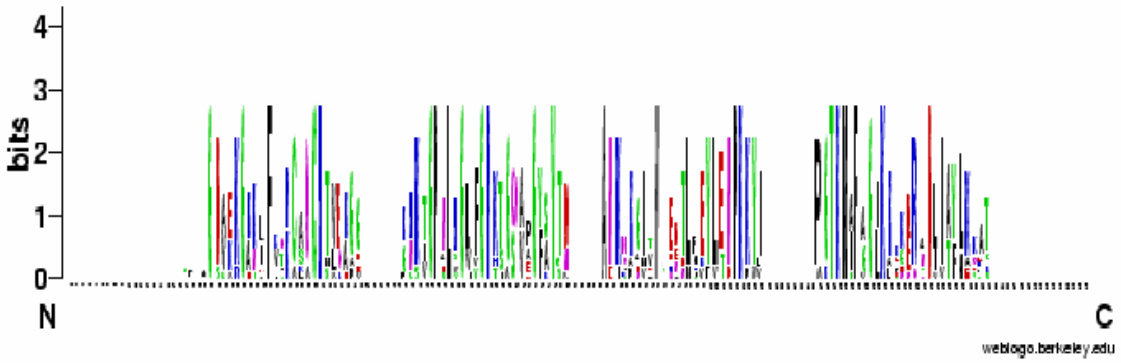
شکل ۱۳-۱۰

```
sp|P00028|CYC_LAMTR      1
----------------------GDVEKGKKVFVQKCSQCHTVEKAG------KHKTGPNL
sp|P62898|CYC_RAT        1
----------------------GDVEKGKKIFVQKCAQCHTVEKGG------KHKTGPNL
sp|P99999|CYC_HUMAN      1
----------------------GDVEKGKKIFIMKCSQCHTVEKGG------KHKTGPNL
sp|P00036|CYC_LUCCU      1
------------------GVPAGDVEKGKKIFVQRCAQCHTVEAGG------KHKVGPNL
sp|P00044|CYC1_YEAST     1
-----------------TEFKAGSAKKGATLFKTRCLQCHTVEKGG------PHKVGPNL
sp|P00048|CYC_NEUCR      1
------------------GFSAGDSKKGANLFKTRCAQCHTLEEGG------GNKIGPAL
sp|Q96VP3|CYC_CURLU      1
-----------------MGFEQGDAKKGANLFKTRCAQCHTLKAGE------GNKIGPEL
sp|P00076|CYC_EUGGR      1
----------------------GDAERGKKLFESRAAQCHSAQKGV-------NSTGPSL
sp|Q00499|CY550_PARVE    1
MKISIYATLAALSLALPAVAQEGDAAKGEKEF-NKCKACHMVQAPDGTDIVKGGKTGPNL
consensus                1                          Gd ekGkkvF  rcaqCHtvekgg
hktGPnL
```

شکل ۱۴-۱۰

LOGOS

Logos وسیله ای بسیار مناسب جهت ایجاد یک نمودار تصویری مختصر و مفید از multiple sequence alignment هاست. شکل شماره ۱۰ – ۱۶، LOGO های مربوط به alignment شکل ۱۰ – ۱ را نشان می دهد.



شکل ۱۵-۱۰

نکات زیر را در نگاه به یک LOGO می توان مشاهده کرد:

- ✓ هر ردیف معادل یک ستون در multiple alignment می باشد.

- ✓ طول هر توالی ای که در alignment مورد نظر، از تعداد بیشتری آمینواسید یکسان تشکیل شده باشد، بیشتر است.(برای توضیح بیشتر: هر alignment ی که تعداد بیشتری از آمینواسیدهای آن از یک نوع باشند طول بزرگتری خواهد داشت و در مقابل آن ردیف هایی که از آمینو اسیدهای متنوع تری تشکیل شده اند طول کوتاه تری را دارا هستند.)
- ✓ اندازه ی هر حرف در هر ستون نشانه ای از تعداد تکرار آن آمینواسید در alignment مربوطه است. مثلا سایز حرف G در توالی که ۲۰ آمینواسید G دارد در مقایسه با توالی که ۱۰ آمینواسید G دارد ۲ برابر خواهد بود.
- ✓ حرفی که در بالاترین نقطه ی هر ردیف قرار می گیرد، بیشترین تکرار را در آن alignment داشته است.

ساخت logo هم برای پروتیین ها و هم برای DNA امکان پذیر است. بدیهی است که هر چه اختلاف بین توالی های یک گروه کمتر باشد logo ی مربوط به آن نیز مفیدتر خواهد بود. یکی از سرور های کامل برای ساخت logo را می توانید در آدرس زیر بیابید:

./"/http://weblogo.berkeley.eduHYPERLINK "http://weblogo.berkeley.edu

این سرور در واقع آسان ترین سرور مورد استفاده می باشد .برای ساخت logo ی مورد نظر، می توانید alignment خود را در محل مشخص شده paste کرده و گزینه ی create logo را بزنید. سرور درخواست شما را انجام خواهد داد.

همچنین سروری با آدرس زیر http://www.cbs.dtu.dk/~gorodkin/appl/plogo.html نیز چنین امکانی را برای شما فراهم می کند منتها با میزان اندکی از پیچیدگی!!!

Logo هنگامی معنی دار خواهد بود که شامل تعداد اندکی از ردیف هایی با conservation بالا باشند. سروری سودمند وجود دارد که هر گروه (alignment)را در multiple alignment شما تشخیص داده و برای آن logo می سازد. این سرور را با آدرس زیر می توانید بیابید: http://blocks.fhcrc.org

برنامه های اینترنتی دیگر برای ویرایش و آنالیز توالی ها

در این بخش سرور هایی برای تنظیم و تصحیح توالی ها، آنالیز و همچنین نمایش آن ها به شما معرفی خواهد شد. البته تعدادی از آن ها در فصول قبلی توضیح داده شده اند و ذکر آن ها در این جا تنها برای یادآوری است. همچنین اگر نیازمند تغییر فرمت توالیهایتان شدید بهتر است بار دیگر به جدول ۲ -۱۰ مراجعه کنید.

چند برنامه برای تنظیم و تصحیح alignment ها

تنها ابزار کامل قابل دسترس آنلاین برای تنظیم alignment ها jalview است. ولی اگر خواستار استفاده از ابزارهایی قدرتمند، اما محدودتر هستید دامنه ی انتخابتان وسیع می باشد. جدول ۵-۱۰ در این زمینه به شما کمک خواهد کرد.

نام سرور	آدرس اینترنتی سرور	توضیحاتی پیرامون سرور
Jalview	www.jalview.org www.es.embnet.org/services http://Molbio/jalview	یک برنامه یJAVA ، و قابل دستیابی در اینترنت.
Kalignview	http://msa.cgb.ki.se	یک برنامه آنلاین مناسب برای نشان دادن alignment.
CINEMA	www.bioinf.man.ac.uk/dbbrowser/CINEMA2.1/	یک برنامه ی Java ی بسیار کامل.
Seaview	http://Pbil.univ-lyonl.fr/software/seaview.html	یک برنامه ی تصحیح کننده ی خوب و با نصب بسیار آسان بر روی کامپیوتر.
Belvu	www.cgr.ki.se/cgr/groups/sonnhammer/Belvu.html	بسیار کارامد برای حذف قسمت های زاید.
Bioedit	www.mbio.ncsu.edu/BioEdit/bioedit.html	بسیار مناسب برای کار با RNA
RALEE	www.sanger.ac.uk/users/sgj/raleee/	یک نمایش گر RNA.
Review	http://bioweb.pasteur.fr/cgi-bin/seganal/review-edital.pl	شامل لیست کاملی از انواع نشانگرهای موجود.

جدول ۵-۱۰: برنامه های ویرایش کننده Multiple sequence alignment

چه سرور هایی در تفسیر Multiple alignment به ما کمک خواهند کرد؟

ظاهر یک alignment بر روی تفسیر دیگران از خود بسیار مؤثر است سرورهایی زیادی که در این زمینه کمک یار شما باشند وجود دارند. برای مثال این سرورها با جدا کردن بخش هایی خاص و یا بیرون کشیدن بلوک هایی از توالی ها، به شما در معنی دار کردن alignment کمک خواهند کرد. در جدول ۶ – ۱۰ لیستی از این منابع در اختیار شما قرار گرفته است.

کدام سرورها امکاناتی برای شکیل کردن Multiple alignment ها را در اختیار ما قرار می دهند؟

Boxshade - که در قسمت قبلی معرفی شد - یکی از ساده ترین ابزارهایی است که امکان تنظیم جنبه های نمایشی alignment ها را برای ما فراهم می کند ولی تنها ارائه دهنده ی این خدمات نیست .جدول ۷ - ۱۰ لیستی از سرورهای گوناگون ارائه دهنده ی این خدمات را در اختیار شما قرار خواهد داد.

نام سرور ارائه دهنده	آدرس اینترنتی سرور مذکور	توضیحاتی پیرامون سرور مذکور
ESPrint	http://espript.ibcp.fr/ESPript/ESPript/	ابزاری بسیار قوی جهت رنگ آمیزی و جلوه های تصویری.
Boxshade	http://www.ch.embnet.org/software/BOX_form.html	رنگ آمیزی با رنگ های سفید و سیاه.
Mview	http://bioweb.pasteur.fr/seqanal/interfaces/mview_blast-simple.html	بر روی BLAST alignment اجرا می کند.

جدول ۶-۱۰: استخراج اطلاعات از یک multiple sequence alignment

ابزار هایی که برای زیباتر کردن یک multiple sequence alignment می توانند به کار بروند

Boxshade که در قسمت های قبلی به آن اشاره شد، یکی از ساده ترین ابزار هایی است که برای زیباتر کردن multiple sequence alignment به کار می رود. جدول ۷-۱۰ چند منبع جایگزین برای این کار را معرفی می کند.

Name	*Address*	*Description*
ESPript	espript.ibcp.fr	A very powerful shading-and-coloring tool
Boxshade	www.ch.embnet.org/software/BOX_form.html	Shading in black and white
Mview	bioweb.pasteur.fr/seqanal/interfaces/mview_blast-simple.html	Can process BLAST alignments

بخش چهارم

یک متخصص شوید!! تکنیک های بیوانفورماتیک پیشرفته

ترجمه: سجاد عباسی

ویرایش: دکتر حسن کربکندی

در این بخش می خوانیم...

این بخش به یک قسمت اصلی از بیوانفورماتیک اختصاص دارد: پیش گویی ها. پیش گویی های صحیح رؤیای زیست شناسان است. این رؤیای خوبی است، چون کار کردن با کامپیوتر ارزان است، ولی کارهای آزمایشگاهی گران هستند.

در این بخش رؤیای نهایی زیست شناسان را به شما معرفی می کنیم: پیش گویی ساختار پروتئین ها. همتای نوکلئیک اسیدی این رؤیا که موفق تر هم بوده، پیش گویی ساختار RNA است. ما به این قسمت "بیوانفورماتیک پیشرفته" می گوییم، چون نتایج این روش ها تنها زمانی مفهوم پیدا می کنند که شما بتوانید آن ها را تفسیر کنید و درک درستی از آن ها داشته باشید. به این روش ها نمی توان با چشم بسته اعتماد کرد. در این بخش به شما می گوییم که چه زمانی لازم است چشمانتان را کاملاً باز نگه دارید!

فصل یازدهم

کار کردن با ساختار های سه بعدی پروتئین ها

در این فصل می خوانیم:

- ✓ بررسی ساختار سه بعدی پروتئین ها و دلیل اهمیت آن ها
- ✓ پیش گویی ساختار سه بعدی پروتئین ها
- ✓ یافتن بهترین مدل سه بعدی برای یک توالی خاص
- ✓ طراحی تصاویری از ساختار های سه بعدی پروتئین جهت استفاده در گزارش ها و سخنرانی ها
- ✓ آنالیز یک توالی از نظر ساختاری
- ✓ طراحی مدل های مولکولی چرخان در کامپیوتر

.....

"از آنجایی که هر ساختار پروتئینی عملکرد منحصر به فرد خود را دارد و توضیح این عملکرد ها از روی ساختار ممکن است، همه پروتئین های ارزش بررسی ساختاری را دارند."

— John C. Kendrew, Nobel Lecture 19 2

معمولاً دانشمندان پروتئین هایی را بررسی می کنند که به عملکرد آن ها علاقمند باشند. امروزه همه زیست شناسان می دانند که رابطه نزدیکی بین ساختار و عملکرد پروتئین ها وجود دارد. زمانیکه یک ویژگی جالب در توالی یک پروتئین دیده می شود، مثلاً قطعات تکراری یا قطعاتی که طی تکامل محافظت شده اند، اولین چیزی که به ذهن می رسد اینست که این ویژگی چه تأثیری بر شکل سه بعدی پروتئین گذاشته است.

در این زمینه سؤالات متداولی به ذهن می رسد:

- ✓ آیا وجود اسید آمینه های خاص در محل های خاصی از توالی در پایداری ساختار پروتئینی نقش دارد؟
- ✓ چرا یک توالی خاص حفاظت شده یا متغیر است؟
- ✓ آیا یک قسمت فرضی از توالی آمینو اسیدی در سطح مولکول پروتئین قرار گرفته است؟
- ✓ آیا یک قسمت فرضی از توالی آمینواسیدی مسقیماً در ایفای نقش پروتئین دخیل است؟
- ✓ آیا یک قسمت خاص از توالی آمینو اسیدی در اتصال مولکول پروتئین به مولکول های دیگر نقش دارد؟

در این فصل شما یاد می گیرید که چگونه ساختار سه بعدی یک توالی خاص را در ذهن خود تصور کنید.

از ساختار اولیه تا ساختار ثانویه

زیست شناسان توالی آمینو اسیدی یک پروتئین را به عنوان "ساختمان اولیه[1]" آن می شناسند، در حالیکه شکل نهایی یک پروتئین، "ساختار سه بعدی" یا "ساختمان سوم" آن نامیده می شود. واضح است که یک ساختار واسطه باید بین دو ساختار مذکور وجود داشته باشد؛ این ساختار، "ساختمان دوم" نام دارد. با انجام اولین مطالعات کریستالوگرافی روی ساختار پروتئین ها مشخص شد که توالی های آمینواسیدی بر اساس قانون خاصی روی هم پیچ می خورند تا مولکولی که دارای فعالیت بیولوژیکی است، به وجود آید.

آمینواسید ها ابتدا با آمینو اسید های کناری خود یک سری نواحی با ساختار منظم و تکرار شونده به وجود می آورند (مثل ستون فقرات). سپس نواحی تشکیل شده روی هم یا روی نواحی که تشکیل ساختار منظم نداده اند، تا می خورند و ساختار سه بعدی به وجود می آید. به این ترتیب آمینواسید هایی که در توالی آمینواسیدی از یکدیگر بسیار دور بوده اند، در ساختار سه بعدی درکنار هم قرار می گیرند.

سه نوع قطعه محلی وجود دارد:

- ✓ هلیکس[2]: که ساختار آن شبیه فنر است و معروف ترین نوع آن α-هلیکس نام دارد.
- ✓ رشته های بسط یافته[3] یا نوع β: آمینواسید ها به صف قرار گرفته اند، و آمینواسید های کنار هم، پشت به هم قرار گرفته اند.
- ✓ حلقه های غیر منظم[4]: زنجیره های آمینو اسیدی نه به صورت هلیکس هستند و نه از نوع β.

پیش گویی ساختار دوم یک توالی آمینواسیدی

پیش گویی ساختار دوم یک توالی آمینو اسیدی از بحث های داغ در دهه ۱۹۹۰ بود. این موضوع حقیقتاً یکی از بزرگترین موفقیت های آن دهه بود. امروزه مراکز اینترنتی خوبی وجود دارند که از مدل های Hidden Markov و شبکه های عصبی برای پیش گویی ساختار دوم پروتئین ها استفاده می کنند. در این بخش نحوه استفاده از PSIPRED، که یکی از بهترین این مراکز است، را به شما نشان می دهیم.

اگر به اندازه کافی پروتئین های مشابه با پروتئین شما در بانک اطلاعاتی مورد استفاده شما وجود داشته باشد، ساختمان دوم پروتئین مورد نظر شما با صحت حدود ۸۰ درصد قابل پیش گویی است. اما توجه داشته باشید که به هر حال این فقط یک پیش گویی است و می تواند کم و بیش غیر صحیح باشد.

برای شروع کار با PSIPRED مراحل زیر را انجام دهید:

۱. پروتئین مورد نظر خود را به صورت یک فایل FASTA.txt در بیاورید.

---

[1] primary structure
[2] helix
[3] extended
[4] random coils

در این مثال از توالی GenPep NP_3 0043, Rickettsia conorii استفاده می کنیم. این پروتئین را می توانید در سایت NCBI (www.ncbi.nlm.nih.gov) پیدا کنید. کمی جلوتر در مبحث "ویژگی های توالی پروتئین در ساختار سه بعدی"، در مرحله ۱ ، نحوه انجام این کار بیشتر توضیح داده شده است.

۲. وارد آدرس اینترنتی bioinf.cs.ucl.ac.uk/psipred/ شوید.

صفحه پیش روی شما، صفحه ی اصلی یک مرکز اینترنتی برای پیشگویی ساختمان پروتئین ها است، که توسط واحد بیوانفورماتیک دانشگاه لندن سرویس دهی می شود.

۳. در پایین صفحه روی گزینه CLICK HERE TO ACCESS THE SERVER کلیک کنید.

۴. توالی مورد نظر خود را در پنجره Input Sequence وارد کنید.

در این قسمت هم فرمت FASTA و هم فرمت ساده (بدون سر عنوان)، قابل قبول هستند.

۵. بخش های Choose Prediction Method و Filtering Options را بدون تغییر باقی بگذارید.

۶. در قسمت Submit Sequence آدرس پست الکترونیک خود را وارد کنید.

در این قسمت همچنین یک فضا برای وارد کردن یک نام کوتاه برای توالی مورد نظر شما وجود دارد، که بعداً می تواند برای به خاطر آوردن موضوع آنالیز شما مفید باشد.

۷. روی دکمه Predict کلیک کنید.

این مرکز اینترنتی نتایج به دست آمده را معمولاً در عرض کمتر از ۳۰ ثانیه برای شما ارسال می کند. در صورتی که تغییری در قسمت Choose Prediction Method ایجاد نکرده باشید، نتایج به دست آمده در واقع پیش گویی ساختمان دوم پروتئین خواهد بود.

شما همچنین می توانید سه روش پیش گویی دیگر را نیز انتخاب کنید، که یکی از آن ها مربوط به بخش های عبور کننده از غشاء[1] پروتئین هاست و دو تای دیگر مربوط به تشخیص تاخوردگی های مولکول است.

اگر همه مراحل به خوبی طی شوند، یک صفحه جدید باز می شود که در آن نوشته شده است که اطلاعات شما برای پیشگویی به درستی ارائه شده است و نتیجه آن به آدرس پست الکترونیکی که در مرحله ۶ وارد کرده اید، فرستاده خواهد شد. همچنین در این صفحه تعداد کارهایی که تا انتهای پروسه باید انجام بدهید، نوشته شده است. تنها کاری که شما در این مرحله می توانید انجام بدهید اینست که منتظر رسیدن جواب از طریق پست الکترونیک باشید. در این مدت می توانید یک قهوه میل کنید!

توجه کنید که این سرور آدرس های پست الکترونیک رایگان مثل Hotmail در MSN یا Gmail در google را قبول نمی کند، بنابراین به این آدرس ها جوابی فرستاده نخواهد شد.

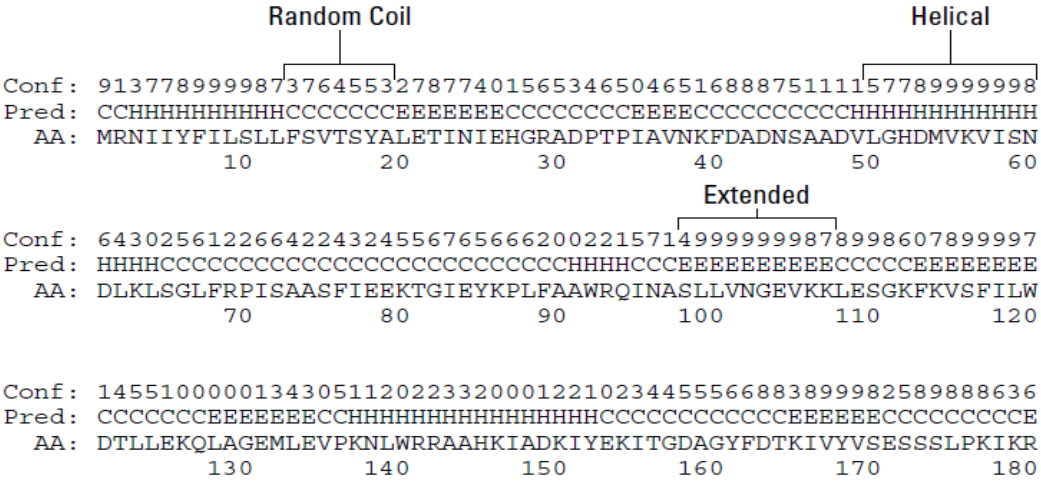
اکنون دنبال یک نامه الکترونیک از psipred@cs.ucl.ac.uk بگردید (یک نمونه آن در شکل ۱-۱۱ نشان داده شده است). نتیجه به صورت سه خط نمایش داده می شود که خط آخر آن (AA) توالی پروتئن مورد نظر شماست. دو خط دیگر عبارتند از:

- ✓ *خط پیش گویی (pred)* که از حروف H، E و C تشکیل شده است و فرم نهایی هر قسمت را نشان می دهد. (H: فرم هلیکس، E: فرم بسط یافته و C: فرم نامنظم)

[1] transmembrane segments

✓ *خط اطمینان (conf)* که از اعداد ۰ تا ۹ تشکیل شده است و میزان قابل اطمینان بودن پیش گویی را در مورد هر جایگاه توالی نشان می دهد (۹ بیشترین و ۰ کمترین اطمینان را نشان می دهد).

برای اینکه بتوانید نتایج مربوط به هر اسید آمینه را بهتر و بدون خطا بخوانید، نتایج را وارد یک فایل Word کنید و از یک فونت مناسب که فاصله مناسبی بین حروف ایجاد می کند (مثلاً فونت Courier 10) استفاده کنید.



شکل ۱-۱۱: نامه الکترونیک فرستاده شده توسط PSIPRED: نتیجه جستجو در مورد R.conorii TolB (NP_3 0043)

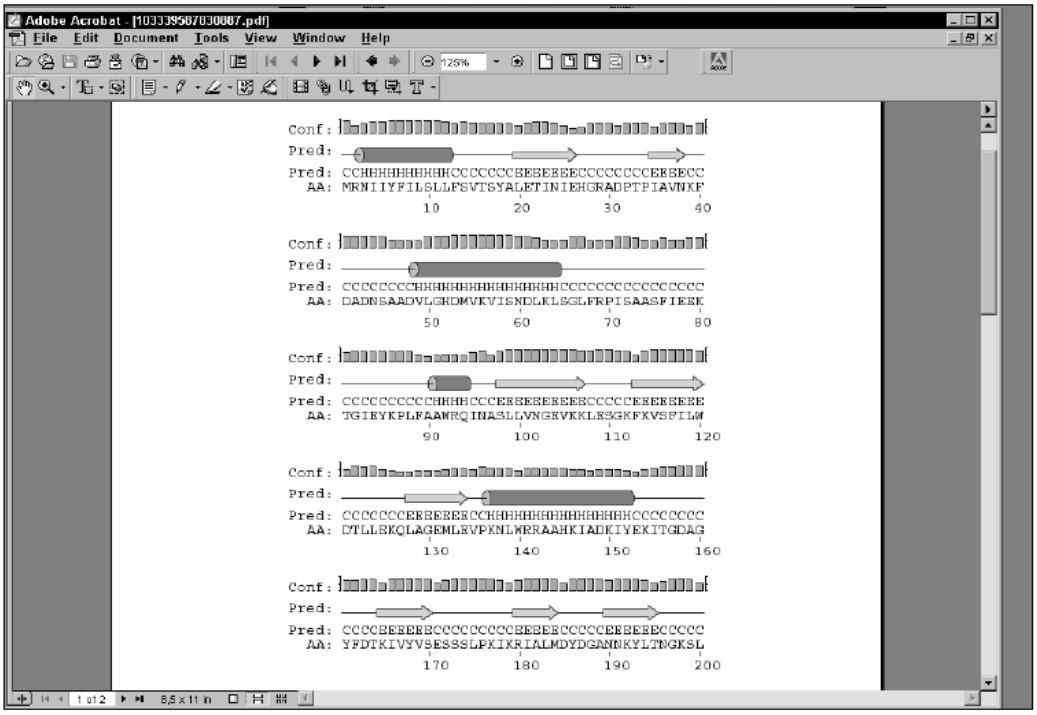
PSIPRED بخاطر صحت بالای نتایج آن (به طور میانگین بالای ۸۰ درصد) و همچنین گرافیک خوب نتایج آن بسیار پر طرفدار است. برای دریافت فایل گرافیکی نتیجه ی به دست آمده به انتهای نامه الکترونیک مراجعه کنید؛ در آن قسمت متن زیر را خواهید دید:

```
Calculate PostScript, PDF and JPEG graphical output for
          this result using
http://bioinf2.cs.ucl.ac.uk/cgi-
          bin/psipred/gra/nphview2.cgi?id=xxxxx.psi
```

با کلیک بر روی لینک مربوطه صفحه ای باز می شود که در آن امکان دنلود فایل های گرافیک در ۴ فرمت متفاوت وجود دارد:

- ✓ **PostScript File:** http://bioinf2.cs.ucl.ac.uk/psiout/xxxxxx.ps
- ✓ **PDF File:** http://bioinf2.cs.ucl.ac.uk/psiout/xxxxxx.pdf
- ✓ **JPEG Page 1:** http://bioinf2.cs.ucl.ac.uk/psiout/xxxxx_1.jpg
- ✓ **JPEG Page 2:** http://bioinf2.cs.ucl.ac.uk/psiout/xxxxx_2.jpg

اگر نرم افزار Adobe Acrobat را روی کامپیوتر خود دارید، می توانید فرمت PDF را انتخاب کنید. در این صورت صفحه ای شبیه شکل ۲-۱۱ باز خواهد شد. اگر می خواهید نتیجه به دست آمده را وارد PowerPoint کنید، بهتر است فرمت JPEG را انتخاب کنید.



شکل ۲-۱۱: صفحه گرافیکی نتایج در PSIPRED: قسمتی از نتیجه پیش گویی مربوط به توالی R.conorii TolB (NP_3 0043)

تفاوت نتایج در این فرمت ها با نتایجی که به فرمت text بودند، در این است که خط اطمینان بـه صـورت نمـودار نـشان داده شده است و در خط پیش گویی فرم هلیکس به شکل سیلندر، فرم بسط یافته به شکل پیکان، و عدم وجود سیلندر یا پیکان نشاندهنده فرم نامنظم است.

پیش گویی اطلاعات بیشتر در مورد خصوصیات ساختاری پروتئین ها

برای رسیدن به پاسخ سؤالات پایه ای مثلاً پیش گویی ساختمان دوم یک پروتئین، PSIPRED یک منبع ایـده آل اسـت. اما برای کسب اطلاعات بیشتر به سایت های دیگری نیز نیاز داریم.

سایت PredictProtein احتمالاً جامع ترین سابت برای آنالیز ساختار پروتئین هاست. متأسفانه سایت اصلی کـه تحـت نظارت دانشگاه کلمبیا است، خیلی مشغول است و ممکن است رسـیدن بـه پاسـخ در ایـن سـایت یـک تـا ۲ روز طـول بکـشد. خوشبختانه این سایت تعداد زیادی سایت های زیر مجموعه در سرتاسر جهان، از جمله اروپـا، ایـالات متحـده، آسـیا و اسـترالیا دارد، اما باید توجه داشته باشید که بعضی از لینک هایی که در این سایت های زیر مجموعه وجود دارند از کار افتاده ومنـسوخ شده اند.

لیستی از لینک های مفید و به روز در این سایت ها به شرح زیر است:

اروپا:

www.predictprotein.org/
www.cmbi.ru.nl/bioinf/predictprotein/

امریکا:

cubic.bioc.columbia.edu/predictprotein/
www.sdsc.edu/predictprotein/

آسیا:

www.cbi.pku.edu.cn/predictprotein/

برای کار کردن با یک سایت PredictProtein توالی پروتئین مورد نظر خود را در یک فایل ذخیره کنید، سپس مراحل زیر را دنبال کنید:

۱. وارد آدرس اینترنتی www.sdsc.edu/predictprotein/ شوید. در حال حاضر سایت San Diego کارا ترین سایت است.
۲. روی لینک Submit a Protein Sequence for Prediction کلیک کنید. یک فرم به صورت پیش نویس ظاهر می شود.
۳. آدرس پست الکترونیک خود را در قسمت بالای فرم وارد کنید.
۴. یک حالت نمایش برای نتایج خود انتخاب کنید.

حالت نمایش پیش نویس بصورت Results on Our Web Site است که در این صورت نتایج در قالب یک نامه الکترونیک به صورت URL برای شما فرستاده می شود.

امکان دسترسی به این فایل تا ۱ روز وجود دارد. این فایل همچنین می تواند به صورت HTML یا HTML for Printout باشد. همچنین تغییرات گرافیکی نیز در این مرحله می تواند اعمال شود.

اگر حالت نمایش پیش نویس را از حالت انتخاب خارج کنید، نتایج به صورت یک فایل خیلی بزرگ برای شما فرستاده می شود که ممکن است سرویس پست الکترونیک مورد استفاده شما قادر به دریافت آن نباشد.

بنابراین در همه حالات، سایت مورد نظر برای شما یک نامه الکترونیک می فرستد، که آن نامه یا خودش حاوی نتایج است، یا شما را به آدرسی که می توانید نتایج را آنجا ببینید راهنمایی می کند.

۵. توالی پروتئین خود را از فایل ذخیره شده وارد پنجره Paste or Type Your Sequence کنید. در این قسمت فقط یک توالی پروتئین، به صورت کدهای یک حرفی و با حالت نمایش ساده قابل قبول است. سر عنوان های FASTA در این قسمت قابل قبول نیستند.
۶. در صورت تمایل می توانید در قسمت One-Line Name of Protein یک نام کوتاه برای توالی خود بنویسید.
۷. روی دکمه Submit/Run Prediction کلیک کنید. تنها کاری که باقی مانده اینست که منتظر رسیدن نامه الکترونیک باشید. برای تخمین زمان این پروسه روی دکمه زرد رنگ Wait در پایین سمت چپ فرم کلیک کنید.

نتایج یک آنالیز معمولی شامل موارد زیر است:

- ✓ پیش گویی ساختمان دوم به صورت حروف H، E و C (H: فرم هلیکس، E: فرم بسط یافته و C: فرم نامنظم)
- ✓ پیش گویی میزان دسترسی حلال یا مایع احاطه کننده مولکول به هر کدام از آمینواسیدها
- ✓ پیش گویی مارپیچ های عبور کننده از غشاء و محل آن ها.
- ✓ پیش گویی مناطق کروی پروتئین
- ✓ پیش گویی مناطق دارای مارپیچ نامنظم پروتئین
- ✓ پیش گویی مناطق تکرار شونده یک توالی (مناطق تکرار شونده در فصل ۶ توضیح داده شده اند)
- ✓ توصیف ساختار فرضی برای نواحی مختلف یک پروتئین (نواحی Prodom)

- ✓ فهرستی از توالی های تقریباً مشابه با توالی مورد نظر شما که با استفاده از برنامه PSI-BLAST در بانک اطلاعاتی Swiss-Prot یافت شده اند.
- ✓ بررسی تطابق توالی مورد نظر شما با توالی های مشابه که با استفاده از برنامه MaxHome انجام می شود.
- ✓ پیش گویی باندهای دی سولفیدی (باندهای سیستئین–سیستئین) یک توالی
- ✓ پیش گویی نواحی غنی از آمینواسیدهای خاص در توالی مورد نظر

از بین این ویژگی ها فقط آنهایی که در توالی شما وجود دارند، در نتایج نشان داده می شوند.

اگر در هر یک از زمینه های ذکر شده به جزئیات بیشتری نیاز دارید، می توانید اطلاعات اولیه را در فرم های Expert یا Advanced وارد کنید. بدین ترتیب برنامه های لازم برای کسب جزئیات بیشتر فعال می شوند.

بخش Metaserver در سایت PredictProtein

سایت PredictProtein حاوی یک بخش به نام Metaserver است که به کاربران این امکان را می دهد که توالی موردنظر خود را در آن واحد به چند سرور ارائه کنند و آن سرور ها به صورت خودکار نتایج را برای کاربران می فرستند.

برای استفاده از این بخش مراحل زیر را دنبال کنید:

۱. وارد آدرس اینترنتی www.sdsc.edu/predictprotein/ شوید.
۲. روی لینک Submit Requests to META-PP کلیک کنید. یک صفحه جدید باز می شود.
۳. آدرس پست الکترونیک خود را در محل مورد نظر وارد کنید. نتایج به آن آدرس فرستاده خوهند شد.
۴. توالی مورد نظر خود را در پنجره Paste or Type Your Sequence وارد کنید. توالی باید با کدهای یک حرفی و با فرمت ساده باشد.
۵. در پایین صفحه روش مورد نظر خود را برای پیش گویی ساختار پروتئین، از جدول مربوطه انتخاب کنید.

اکثر این سرور ها از راه های فرعی برای پیش گویی ساختار پروتئین ها یا ایجاد ارتباط بین توالی مورد نظر شما با یک توالی شناخته شده (threading method) استفاده می کنند.

۶. روی دکمه Submit/Run Prediction کلیک کنید.

اگر چه ارائه کردن چند آنالیز در یک زمان ممکن است کار راحتی به نظر برسد، اما تجربه ما نشان داده است که انجام آنالیز ها به صورت مستقیم در هر کدام از سایت های مربوطه، سریع تر و راحت تر از زمانی است که این کار را با واسطه Metaserver انجام می دهید.

یافتن بهترین سرور

سایت های زیادی در سرتاسر جهان وجود دارند که هر کدام الگوریتم خاص خود را دارند و به همین دلیل بعضی سریع تر از بعضی دیگر عمل می کنند. برای اینکه بهترین سایت را پیدا کنیم، به یک سیستم ناظر بر سرور های پیش گویی ساختمان دوم پروتئین ها، به نام EVA [1] و به آدرس cubic.bioc.columbia.edu/eva/ مراجعه کنید.

[1] *EV*aluation of *A*utomatic protein structure prediction

از ساختار اولیه تا ساختار سه بعدی

در قسمت قبل آموختید که چگونه اطلاعات پایه ای در مورد پیش گویی ساختار پروتئین دارای فعالیت بیولوژیک از روی توالی آمینواسیدی آن به دست آورید. این اطلاعات شامل نواحی هلیکس و بسط یافته، حلقه ها، آمینواسیدهایی که در سطح مولکول و در معرض محیط بیرون قرار می گیرند، و ... می باشد.

زمانیکه شما هیچ اطلاعاتی ندارید، این اطلاعات اولیه می توانند خیلی مفید باشند. مثلاً این اطلاعات در پیش گویی اثرات احتمالی یک جهش ژنی یا انتخاب قطعه پروتئینی مناسب برای ساخت یک آنتی بادی می تواند مفید باشد. اما متأسفانه اطلاعاتی که پیش گویی ساختار دوم پروتئین ها به ما می دهد، بسیار کمتر از اطلاعات و جزئیات لازم برای توصیف مولکول واقعی است. در این قسمت یاد می گیرید که چگونه برای درک بهتر وقایعی که روی توالی آمینواسیدی اتفاق می افتد، اطلاعاتی در مورد ساختار سه بعدی جمع آوری کرده و چگونه از این اطلاعات استفاده کنید.

خوشبختانه اطلاعات تجربی زیادی در مورد ساختار سه بعدی پروتئین ها در اینترنت در دسترس است. همانگونه که زیست شناسان به توافق رسیده اند که اطلاعات موجود در مورد توالی ها را در بانک اطلاعاتی GenBank/EMBL/DDBJ جمع آوری کنند، این توافق نیز حاصل شد که اطلاعات موجود در مورد ساختارهای سه بعدی، در یک بانک اطلاعاتی به نام Protein Data Bank جمع آوری شود. این بانک اطلاعاتی بیشتر با نام اختصاری PBD شاناخته می شود.

یافتن و نمایش یک ساختار سه بعدی

قبل از اینکه جستجو در Protein Data Bank (PDB) را شروع کنید، سعی کنید اطلاعات دقیقی در مورد ساختاری که به دنبال آن می گردید، داشته باشید؛ مثلاً نام دقیق پروتئین و شماره شناسایی آن. شما می توانید شماره شناسایی از منابعی مثل ExPASy/Swiss-Prot یا با استفاده از ابزار های جستجوی NCBI به دست آورید (توضیح بیشتر در این مورد در فصل ۴).

در این قسمت یاد می گیرید که چگونه یک ساختار را در PBD به دست آورده و نمایش بدهید. به عنوان مثال فرض کنید می خواهید ساختار یکی از پروتئین های باکتری *Escherichia coli (E. coli)* به نام TolB و با کد شناسایی 1CRZ را در PBD جستجو کنید:

۱. وارد آدرس اینترنتی www.rcsb.org/pdb/ شوید.
   بدین ترتیب وارد صفحه اصلی سایت PBD می شوید.
۲. در پنجره بالای صفحه کد شناسایی را وارد کنید.
۳. روی گزینه SEARCH کلیک کنید.

صفحه ای مشابه شکل ۳-۱۱ در مقابل شما ظاهر خواهد شد.این صفحه که فرم Structure Summary نامیده می شود، حاوی اطلاعات ضروری در مورد پروتئین مورد نظر، از جمله یک شکل کوچک از آن، یک منبع مرتبط، عملکرد و نقش آن، ارگانیسم منشأ آن و احتمالاً اطلاعات اضافی در مورد بلورشناسی می باشد. با کلیک کردن بر روی گزینه هایی که در انتهای فرم وجود دارند، می توانید جزئیات بیشتری در مورد جنبه های مختلف این پروتئین به دست آورید (شکل ۳-۱۱). اکنون هدف ما فقط اینست که شکل سه بعدی پروتئین را ببینیم.

۴. در قسمت Display Options روی لینک All Images کلیک کنید.

صفحه جدیدی باز می شود که در بالای آن شکل بزرگتری از ساختار پروتئین و در پایین آن یک شکل متحرک (یک برنامه Java) وجود دارد.

۵. روی گزینه رنگی Still Image راست-کلیک کنید.

یک منو باز می شود که امکان پرینت کردن، ذخیره کردن یا کپی کردن این فایل را به شما می دهد. این فایل اکنون برای وارد کردن به گزارشات یا فایل های سخنرانی شما مناسب است (چنین تصویر زیبایی می تواند رئیس یا استاد شما را تحت تأثیر قرار دهد!).

اما صبر کنید! با داشتن یک تصویر متحرک از ساختار پروتئین می توانید بهتر آن ساختار را درک کنید. برای این کار ادامه مراحل را دنبال کنید:

۶. روی تصویر متحرک راست کلیک کنید.

یک منو بزرگ از گزینه های مختلف باز می شود. آن ها را باز کنید تا ببینید کدامیک با سیستم کامپیوتر شما قابل نمایش است. به عنوان مثال با انتخاب گزینه Spin On مولکول به طور دائم در حال چرخش خواهد بود (برای متوقف کردن آن کافی است روی گزینه Spin Off کلیک کنید).

۷. اگر چپ-کلیک را روی هر قسمت از تصویر پروتئین نگه دارید، می توانید شکل را به هر جایی از صفحه منتقل کنید (شکل ۴-۱۱)، و همچنین به صورت کنترل شده آن را بچرخانید. بدین ترتیب می توانید نواحی خاص پروتئین را از هر زاویه ای که بخواهید، بررسی کنید. همچنین می توانید یک قسمت ساختار را به صورت بزرگ شده ببینید. با انتخاب یک محل خاص از ساختار پروتئین ، می توان فواصل یا زاویه های مورد نظر را اندازه گرفت. این قسمت خیلی جالب است، حتماً امتحان کنید!

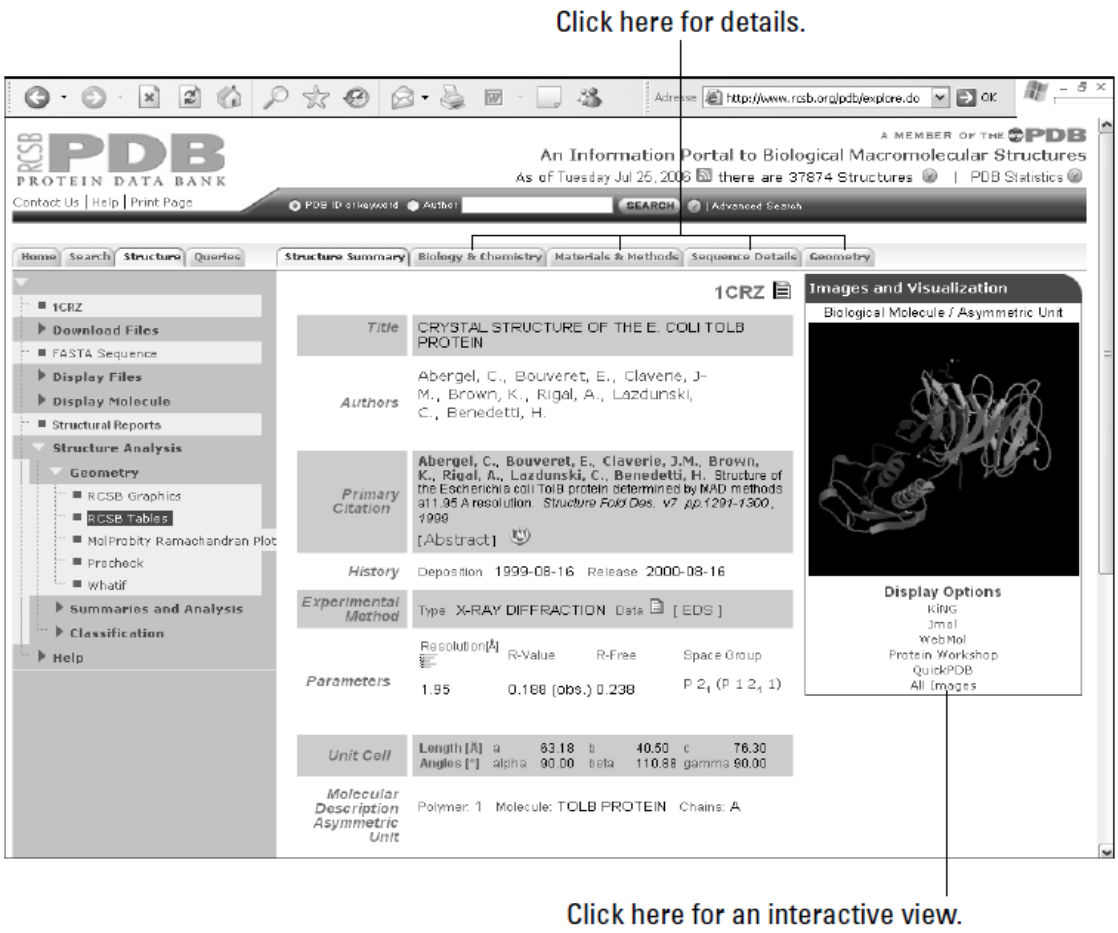
سعی کنید گزینه های دیگر را در Display Options امتحان کنید (... ,KiNG, WebMol)، و ببینید که کدامیک از نظر شما مناسب ترند. بخش "بررسی ارتباط بین توالی و ساختار PDB با استفاده از تصاویر متحرک" را در ادامه این فصل ببینید. در این بخش اطلاعاتی در مورد یک ابزار پیشرفته تر ارائه شده است، این ابزار نیاز به دانلود کردن و نصب نرم افزار CnD3 روی کامپیوتر دارد.

در مراحل فوق به شما نشان دادیم که چگونه می توان از یک وروری PBD مستقیماً به تصویر گرافیکی از یک مولکول پروتئین رسید. اما ممکن است شما بخواهید ورودی PBD را، برای انجام مطالعات بیشتر، روی کامپیوتر خود ذخیره کنید. کمی جلوتر چگونگی انجام این کار را یاد می گیرید.

افراد غیر متخصص قادر به خواندن فایل های PDB نیستند. این فایل ها بسیار بزرگ و شامل مقدار بسیار زیادی داده هستند. به عنوان مثال فایل 1CRZ حاوی مختصات فضایی x، y و z مربوط به ۳۴۸۸ اتم و همچنین اطلاعات اضافی در مورد نحوه اتصال آن هاست.

اگر می خواهید فایل PBD را دانلود کنید، باید در مرحله ۳ به نوار آبی رنگ سمت چپ صفحه توجه کنید (شکل ۳-۱۱). در این نوار چند گزینه وجود دارد:

- ✓ Download Files: برای ذخیره کردن اطلاعات در کامپیوتر
- ✓ Get the FASTA Sequence،
- ✓ Display Files،



شکل ۳-۱۱: نتیجه جستجوی ساختار 1CRZ در PBD

- ✓ Display Molecule: برای فهمیدن فرمت داخلی آن فایل ها
- ✓ Structural Reports،
- ✓ Structure Analysis،
- ✓ Help

روی لینک اول (Download Files) کلیک کنید تا مختصات به دست آمده از اتم های مولکول را دانلود و ذخیره کنید. به همین ترتیب گزینه های دیگر و همچنین گزینه های موجود در نوار بالای صفحه را امتحان کنید. مثلاً با کلیک روی لینک Sequence Details یک شکل زیبا از ساختمان دوم واقعی مولکول همراه با توالی آمینواسیدی آن ظاهر خواهد شد. حال می توانید این شکل را با شکل های پیش بینی شده که در ابتدای فصل به دست آورده بودید، مقایسه کنید.

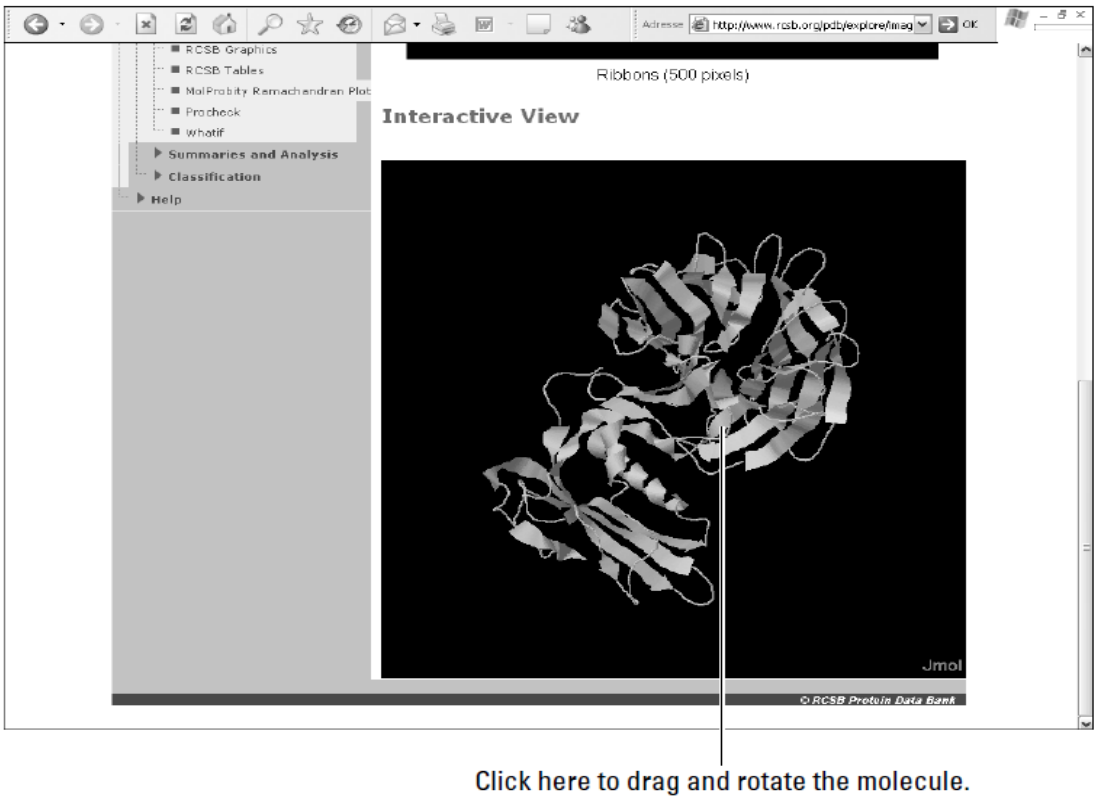
حدس زدن ساختار سه بعدی یک پروتئین

در بخش قبل یاد گرفتید که چگونه یک ساختار سه بعدی شناخته شده که قبلاً در PBD ثبت شده است را بازیابی کنید و نمایش بدهید. این کار قطعاً بسیار مفید است، اما همه اهدافی را که در ابتدای این فصل تعیین کردیم، تأمین نمی کند. در واقع ما هنوز نمی دانیم قطعات جالب توالی ما در ساختار سه بعدی چگونه به نظر می رسند.

در ادامه این فصل با دنبال کردن یک سری مراحل، یاد می گیرید که چگونه می توانید گاهی فقط با داشتن توالی آمینواسیدی یک پروتئین و انجام یک عملیات ساده، به دید نسبتاً خوبی در مورد ساختار سه بعدی پروتئین مورد نظر برسید.

ما فرض می کنیم که شما توالی و ساختمان دوم پروتئین خود را تعیین کرده اید و اکنون به دنبال پاسخ این سؤال هستید که: "چگونه این ساختمان روی هم تا می خورد وشکل سه بعدی مولکول را به وجود می آورد؟" یک راه ساده برای رسیدن به جواب این سؤال اینست که در PBD دنبال یک پروتئین مشابه و با ساختار سه بعدی شناخته شده بگردید. برای این کار مراحل زیر را دنبال کنید.

فرض کنید توالی آمینواسیدی مربوط به ژن TolB باکتری *Rickettsia conorii* را تعیین کرده اید، و به دنبال ساختار پروتئین مرتبط با آن می گردید:



شکل ۴-۱۱: نمای Interactive از ورودی 1CRZ بانک اطلاعاتی PBD

۱. توالی آمینواسیدی پروتئین را از سایت NCBI پیدا کنید:

أ- وارد آدرس اینترنتی www.ncbi.nlm.nih.gov شوید.

ب- با کلیک روی گزینه Search یک منو باز می شود. گزینه Protein را انتخاب کنید.

ت- در پنجره جستجو کد شناسایی NP_3 0043 را وارد کرده روی گزینه Go کلیک کنید.

ث- پس از ظاهر شدن نتیجه ی جستجو فرمت نمایش را به FASTA تغییر دهید.

توالی پروتئین TolB مربوط به *Rickettsia conorii* اکنون آماده استفاده است.

۲. یک صفحه اینترنت جدید باز کنید و وارد سرور BLAST NCBI به آدرس اینترنتی www.ncbi.nlm.nih.gov/BLAST/ شوید.

۳. روی لینک Standard Protein-Protein BLAST [blastp] کلیک کنید.

این لینک در بالای سمت راست صفحه قرار گرفته است. صفحه ی اصلی BLAST ظاهر خواهد شد.

۴. گزینه pdb را از منوی Choose database انتخاب کنید.

۵. توالی به دست آمده از پروتئین TolB باکتری *Rickettsia conorii* را وارد پنجره جستجو کنید.
۶. برای ساده تر کردن نتایج مربع کنار گزینه Do CD-search را از حالت انتخاب خارج کنید.
۷. روی گزینه BLAST کلیک کنید.

یک صفحه جدید باز می شود که در آن نوشته شده است که درخواست شما با موفقیت توسط سرور دریافت شده است.

۸. در صفحه جدید روی گزینه Format کلیک کنید و منتظر بمانید تا جستجو انجام شود.

شکل ۵-۱۱ نمونه ای از انواع نتایجی که مورد انتظار است را نشان می دهد. در دو مورد از این نتایج عدد E[1] بسیار بزرگتر از ۱ است. این دو نتیجه، با کدهای شناسایی 1C5K و 1CRZ مشخص شده اند. اگر روی آن ها کلیک کنید، خواهید دید که این دو مربوط به پروتئین TolB باکتری *E.coli* هستند که ساختار آن ها به صورت جداگانه در دو آزمایشگاه مختلف تعیین شده است.

به دو دلیل این نتیجه خیلی خوب است:

۱. توالی های شناخته شده مشابه ثبت شده در PBD با توالی مورد نظر ما ۲۹ درصد شباهت دارند،
۲. نواحی مشابه در سراسر توالی وجود دارند.

تجربه نشان می دهد که با چنین درصد تشابهی، ساختار سه بعدی پروتئین مورد نظر ما (TolB باکتری Rickettsia conorii) احتمالاً ظاهری شبیه به پروتئین معادل آن در E. coli دارد. بدین ترتیب ما یک مدل تقریباً کامل داریم که می توانیم روی آن کار کنیم. در بخش بعد یاد می گیرید که چگونه با استفاده از این مدل تفسیر بیشتری روی توالی پروتئین TolB داشته باشید.

بررسی ویژگی های یک توالی در شکل سه بعدی

از یک توالی به تنهایی می توان اطلاعات زیادی به دست آورد، اما با مقایسه آن با توالی های دیگر اطلاعات بیشتر و ارزشمند تری به دست می آید. در فصل ۹ با تکیه بر تطابق و مقایسه پروتئین های مختلف سعی کرده ایم مهمترین نواحی توالی ها را مشخص کنیم: جایگاه های محافظت شده و جایگاه های بسیار متغیر توالی ها برای پیش گویی و فهم نقش و عملکرد یک پروتئین مهم هستند. بعلاوه با مشخص کردن محل فضایی دقیق جایگاه های محافظت شده ی توالی، می توان اطلاعات بیشتری در مورد نقش بیولوژیک پروتئین ها به دست آورد؛ به عنوان مثال این جایگاه ها می توانند حفره ای باشند که محل فعال یک آنزیم را تشکیل می دهند، یا اگر این جایگاه ها در سطح مولکول باشند، می توان فرض کرد که محل هایی برای بر هم کنش با مولکول های دیگر هستند، و مثال هایی از این قبیل.

با استفاده از مثال خانواده پروتئین TolB (که نقش آن ها نسبتاً ناشناخته است)، می توان اثر متقابل تطابق مولکول های پروتئین و آنالیز ساختاری را نشان داد. برای این منظور مراحل زیر را دنبال کنید:

۱. ابتدا چند توالی مشابه TolBاز گونه های مختلف باکتریایی را در NCBI پیدا کنید:
   أ- وارد آدرس اینترنتی www.ncbi.nlm. nih.gov شوید.
   ب- در منوی Search گزینه Protein را انتخاب کنید.

[1] E-value

ت- در پنجره جستجو کدهای شناسایی (R. conorii) NP_3 0043، (E. coli) NP_4152، (Y. pestis) NP_404737، (P. aeruginosa) NP_249 3 و (H. influenzae) NP_43 543 را وارد کنید، سپس روی گزینه Go کلیک کنید.

ث- زمانیکه نتیجه ی جستجو ظاهر شد، فرمت نمایش آن را به FASTA تغییر دهید.

ج- در نهایت روی گزینه Send to Text کلیک کنید تا حروف اضافی حذف شوند.

پنج توالی پروتئین TolB اکنون آماده استفاده هستند و می توان تطابق بین آن ها را بررسی کرد.

۲. در یک صفحه جدید وارد آدرس اینترنتی www.igs.cnrs-mrs.fr/Tcoffee/ شوید.

۳. در ابتدای خط TCOFFEE option روی گزینه Regular کلیک کنید.

۴. پنج توالی مربوط به TolB را وارد پنجره Tcoffee کنید.
حتماً سرعنوان های فرمت FASTA را هم وارد کنید.

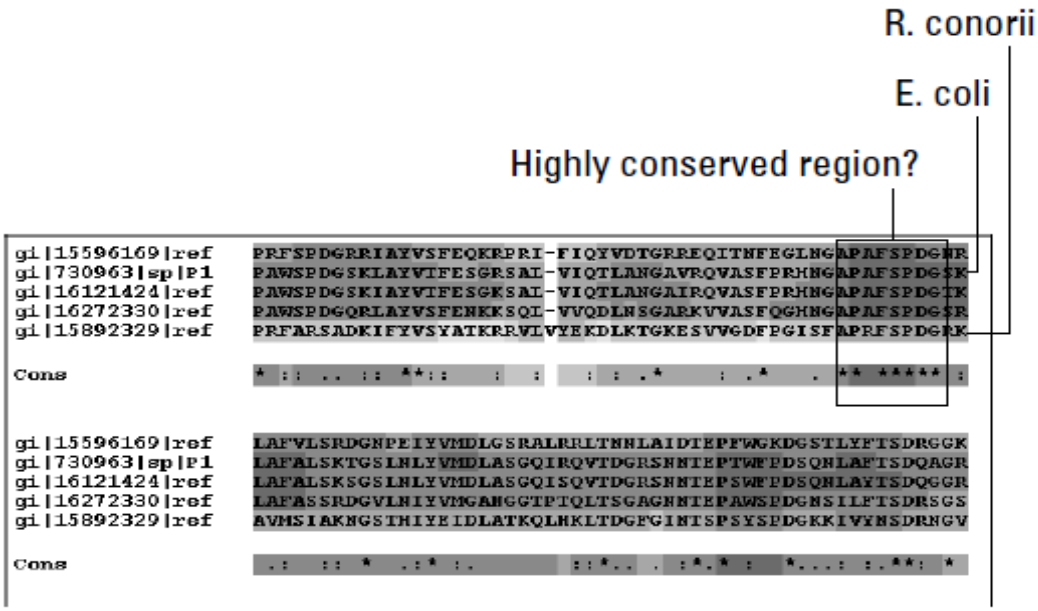
۵. روی گزینه Submit در انتهای فرم کلیک کنید.
فرم Tcoffee Results ظاهر می شود.

۶. از منوی Multiple Alignment Output گزینه score_html را انتخاب کنید.

تطابق توالی ها با رنگ های مختلفی نشان داده می شود؛ قابل اعتماد ترین تطابق با رنگ قرمز مشخص می شود (شکل ۶-۱۱). همانطور که در شکل می بینید، یک ناحیه بسیار حفاظت شده مشترک در تمام توالی های مربوط به TolB وجود دارد. در ادامه این فصل خواهید دید که این نواحی در کدام قسمت ساختار TolB قرار گرفته اند.

بررسی ارتباط بین توالی و ساختار PDB با استفاده از تصاویر متحرک

برای تفسیر اهمیت یک قطعه بدون تغییر یا محافظت شده در پروتئین های TolB، لازم است که ساختار سه بعدی چرخان آن را داشته باشیم، تا بتوانیم آن را از همه زوایا بررسی کنیم. همچنین به طور همزمان باید قادر باشیم به توالی پروتئین رجوع کنیم تا جایگاه های جالب توجه توالی را مشخص کنیم.



شکل ۶-۱۱: یک ناحیه بسیار محافظت شده در پروتئین TolB

اگر چه تا مدتی قبل این کار فقط توسط بلورشناسان و در کامپیوتر های خیلی گران قیمت ممکن بوده است، اگر تا آخر این فصل همراه ما باشید، شما نیز خواهید توانست این کار را در کامپیوتر شخصی خود، در عرض چند دقیقه انجام دهید:

۱. وارد آدرس اینترنتی www.ncbi.nlm.nih.gov/Structure/ شوید.

۲. در پنجره Search Entrez Structure/MMDB در بالای صفحه، کد PDB مدل مورد نظر خود را وارد کنید. برای مثالی که دنبال می کردیم، کد 1CRZ را وارد می کنیم.

اگر چندین گزینه برای انتخاب ساختار وجود داشته باشد، ساختاری را انتخاب کنید که بیشترین قدرت تفکیک[1] را دارد. در مورد مثال ما، بهترین قدرت تفکیک را 1CRZ دارد (۱/۹۵ آنگستروم). عدد آنگستروم هر مدل در بخش کارت ID در ورودی بانک های اطلاعاتی پروتئین ها وجود دارد (شکل ۳-۱۱).

۳. روی دکمه Go کلیک کنید. یک صفحه جدید باز می شود.

۴. روی لینک 1CRZ کلیک کنید.

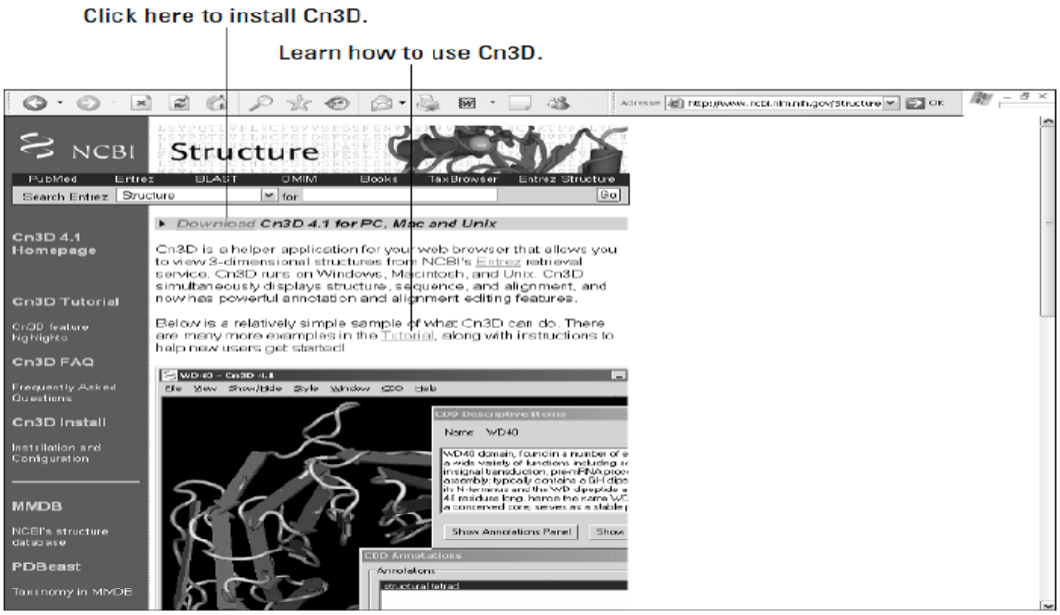
اکنون شما می توانید کارت ساختار MMDB را ببینید (شکل ۷-۱۱).

۵. روی دکمه View 3-D Structure کلیک کنید.

با این کار فایلی دانلود می شود که حاوی مختصات اتم های تشکیل دهنده ی ساختار پروتئین 1CRZ است. این فایل با برنامه Cn3D خوانده می شود. اگر تاکنون با کامپیوتر خود تصویر سه بعدی یک پروتئین را ندیده باشید، صفحه ای باز می شود با این پیام که "با این فایل چه کار می خواهید بکنید؟" این پیام بدین معنی است که کامپیوتر شما فاقد برنامه Cn3D است، بنابراین برای ادامه کار باید این برنامه را روی کامپیوتر خود نصب کنید. بدین منظور به صورت زیر عمل کنید:

أ- روی گزینه Download Cn3D! کلیک کنید.

صفحه اصلی برنامه Cn3D باز می شود (شکل ۸-۱۱). در این قسمت می توانید با کلیک بر روی گزینه Tutorial راهنمای استفاده از برنامه Cn3D را ببینید.



شکل ۸-۱۱: صفحه اصلی Cn3D

[1] resolution

ب- روی لینک Download Cn3D for PC, Mac, or Unix کلیک کنید.

ت- در پنجره ای که باز می شود (شکل ۹-۱۱)، نسخه ای از برنامه را که برای کامپیوتر شما مناسب است، انتخاب کنید: PC، Mac یا Unix.

هر کدام را که انتخاب کنید، یک صفحه جدید با یک آدرس FTP و مشابه با صفحه ای که در شکل ۱۰-۱۱ نشان داده شده، باز می شود.

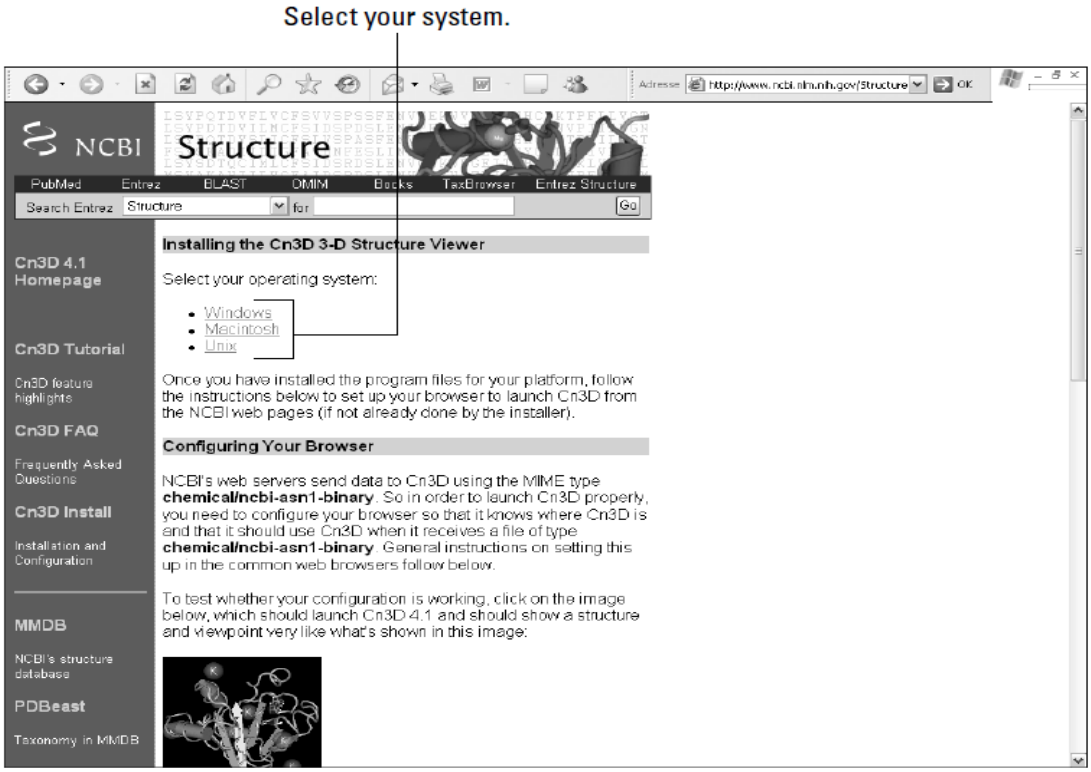
ث- روی لینک آدرس FTP کلیک کنید.

برای سیستم ویندوز ما لینک مورد نظر به صورت زیر است:

ftp://ftp.ncbi.nih.gov/cn3d/Cn3D-4.1.msi

با این کار برنامه Cn3D را دانلود می شود.

ممکن است سیستم از شما بخواهد که نسخه جدید تری از Windows Installer را نصب کنید. شما فقط دستوالعملی که ویندوز به شما می دهد را انجام دهید.



شکل **۹-۱۱**: نصب برنامه Cn3D و انتخاب سیستم اپراتور مناسب برای نصب

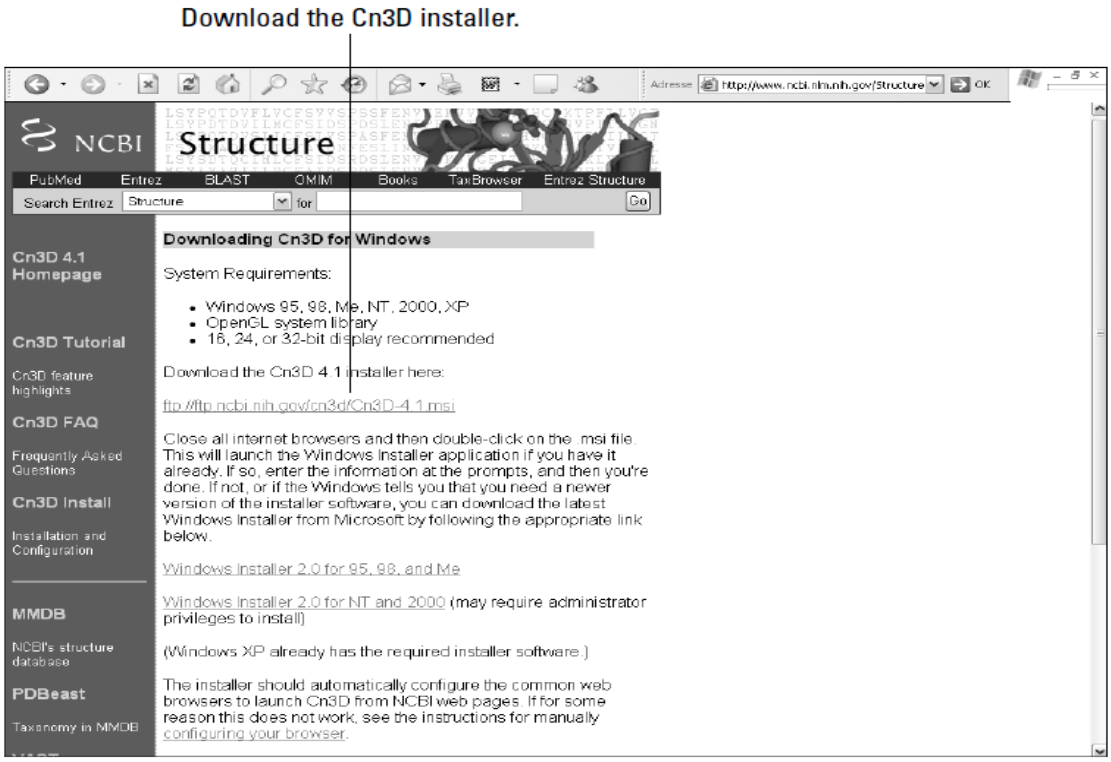
ج- پس از انجام این مراحل، فایل Cn3D-4.1.msi را روی صفحه اصلی کامپیوتر خود منتقل کنید، سپس روی آن دو بار کلیک کنید.

ح- دستورالعمل های نصب برنامه را دنبال کنید.

اکنون که برنامه مورد نیاز برای دیدن ساختار سه بعدی پروتئین را نصب کرده اید، به سایت NCBI برگردید و مراحلی را که در قسمت قبل ذکر شد، تکرار کنید.

۶. روی دکمه View 3D Structure دو بار کلیک کنید.

صفحه ای شبیه شکل ۱۱-۱۱ باز می شود. در بالای صفحه ی MMDB، ساختار سه بعدی پروتئین که مربوط به ورودی 1CRZ بانک اطلاعاتی PBD است، نشان داده شده است.



شکل ۱۰-۱۱: نصب برنامه Cn3D: نصب کننده برنامه روی کامپیوتر

۷. پروتئین خود را بچرخانید.

این کار را به سادگی با حرکت دادن موس[1] کامپیوتر خود می توانید انجام بدهید.

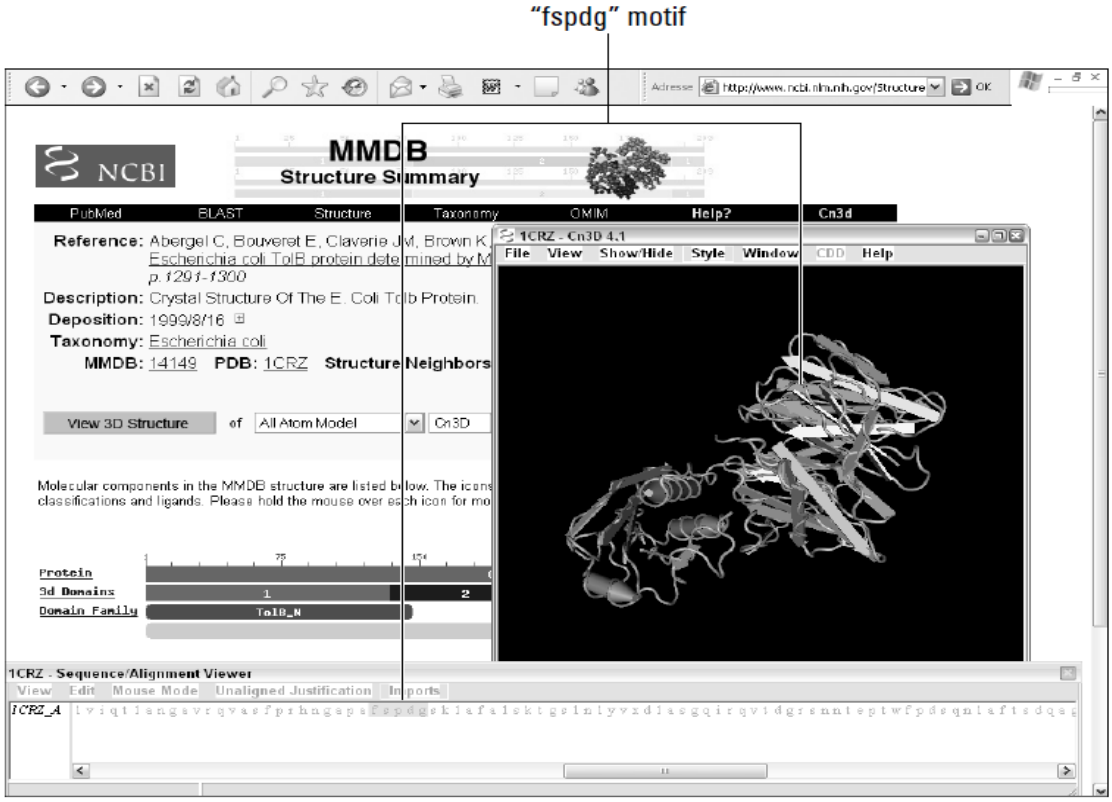
۸. پروتئین خود را برای گرفتن یک عکس آماده کنید.

با استفاده از منو های بالای صفحه ی عکس می توانید بزرگنمایی تصویر را کم یا زیاد کنید. همچنین می توانیدظاهر مدل خود را عوض کنید (wire، tube، ball and stick و حتی solid space filling).

۹. با انتخاب گزینه Export PNG از موی File یک تصویر از پروتئین خود ذخیره کنید.

این تصاویر تماشایی و زیبا به طور مستقل توسط کامپیوتر شما نشان داده می شود و نیازی به اتصال به اینترنت ندارد.

[1] mouse



شکل ۱۱-۱۱: تصویر سه بعدی متحرک از توالی و ساختار سه بعدی پروتئین TolB باکتری *E. coli*

آنالیز ارتباط بین توالی و ساختار با استفاده از اشکال متحرک

ممکن است پنجره دیگری در پایین مولکول متحرک دیده شود (شکل ۱۱-۱۱). این پنجره sequence viewer نام دارد. این پنجره بصورت پیش نویس، توالی پروتئین مرتبط با ورودی PBD را نشان می دهد. مثلاً ما توالی مولکول TolB باکتری *E. coli* را در آن می بینیم.

اکنون زمان آنست که در توالی به دنبال جایگاه های بدون تغییر، که به صورت FSPDG خوانده می شوند، بگردیم:

۱. در نوار بالای پنجره توالی روی دکمه view کلیک کنید.
۲. از منوی حاصل گزینه Find Pattern را انتخاب کنید.
۳. یک پنجره باز می شود، در آن FSPDG را تایپ کنید.

قطعه مورد نظر هم در توالی و هم در شکل پروتئین به رنگ زرد در می آید. با حرکت در امتداد توالی با چرخاندن مولکول در شکل می توانید این قسمت زرد رنگ را مشاهده کنید.

با کلیک بر روی یک جایگاه در توالی یا یک قطعه از آن در قسمت sequence viewer، می توانید آن را به حالت انتخاب در بیاورید.

اما این قطعات بدون تغییر چه فایده ای برای ما دارند؟ همانطور که در شکل ۱۱-۱۱ می بینید، قطعه بدون تغییر توالی ما در یک حلقه که کاملاً در سطح مولکول است، واقع شده است.حلقه های واقع شده در سطح مولکول پروتئین معمولاً متغیرترین بخش های مولکول هستند، چون بیشتر در معرض تغییرند. بنابراین اگر این قطعات محافظت شده و بدون تغییر باشند،

می توانند نشاندهنده نقش مهم آن ها در عملکرد پروتئین، مثلاً شرکت در بر هم کنش با پروتئین ها یا مولکول های کوچک دیگر باشد.

فراتر از این فصل

تاکنون در این فصل یاد گرفتید که یک زیست شناس چگونه با استفاده از کامپیوتر، ساختار پروتئین ها را آنالیز می کند. مطالب این فصل تقریباً ۹۰ درصد اطلاعات لازم جهت بررسی ارتباط بین توالی ها و ساختار های سه بعدی احتمالی آن ها توسط یک فرد غیر حرفه ای را در اختیار شما قرار داده است. در واقع روش های پیچیده و ابزار های دیگری برای این منظور وجود دارند که در این فصل ذکر نشده اند. اما در ادامه فصل سایر روش ها و منابع مرتبط با آن ها را مختصراً معرفی خواهیم کرد:

پیدا کردن پروتئین های با ساختار مشابه

پس از تعیین ساختار سه بعدی یک پروتئین، ممکن است به دنبال پاسخ این سؤال بگردید که: "آیا این پروتئین از نظر شکل ظاهری، بی نظیر است[1]، یا پروتئین های مشابه آن هم وجود دارند؟"

بدین منظور یک فایل PBD حاوی مختصات اتم های تشکیل دهنده پروتئین مورد نظر درست کنید و آن را در قسمت NCBI's structure-structure similarity search service (VAST) سایت NCBI به آدرس www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html وارد کنید.

سایر نمایشگر های PBD

غیر از برنامه Cn3D برنامه های قدرتمند زیادی در اینترنت در دسترس هستند، مانند:

- ✓ **RasMol:** www.rasmol.org
- ✓ **DeepView, Swiss-PdbViewer:** swissmodel.expasy.org/spdbv/

دسته بندی ساختار های پروتئینی PBD

اکنون دسته بندی ساختار های پروتئینی بر اساس نوع تاخوردگی و ریخت شناسی دامین های آن یک بحث بسیار داغ است. مشکل این کار، شناسایی و کمی کردن تشابهات ساختاری بین پروتئین های با خویشاوندی خیلی دور و همچنین مشکل threading است (بخش "بررسی نحوه تاخوردگی توالی ها و تشکیل ساختار های PDB" را ببینید.). سه سایت اصلی در این زمینه عبارتند از:

- ✓ **The CATH classification:** www.cathdb.info
- ✓ **The Dali 3D neighbor finding server:** www.ebi.ac.uk/dali/
- ✓ **The SCOP classification:** scop.mrc-lmb.cam.ac.uk/scop/

[1] به چنین پروتئین هایی در اصطلاح new fold می گویند.

مدل سازی از روی یک پروتئین مشابه

مدل سازی دقیق ساختار سه بعدی یک پروتئین که توالی آن تشابه زیادی با یک پروتئین با ساختار شناخته شده دارد، در واقع صرفه جویی در انرژی و وقت است. در این تکنیک تنها کافی است آمینواسیدهای متفاوت با الگوی اصلی را اضافه، و بر اساس آن تغییرات شکل فضایی را ایجاد نمود.

دو منبع مناسب در این زمنیه عبارتند از:

- ✓ **Modeller:** salilab.org/modeller/modeller.html
- ✓ **SWISS-MODEL:** swissmodel.expasy.org

مدل سازی یک پروتئین از روی توالی آمینواسیدی آن در کامپیوتر

بعضی دانشمندان در تلاشند با شبیه سازی فرایند تا خوردن توالی آمینواسیدی، این امکان را به وجود آورند که تنها با داشتن توالی آمینواسیدی مدل ساختاری پروتئین را به دست آورند. به این فرایند تا خوردن in silico یا ab initio می گویند. امکان استفاده از این روش تاکنون برای توالی های کوچک و خاصی فراهم آمده است. برای آشنایی بیشتر با این زمینه به سایت Folding@home project به آدرس folding. stanford.edu/ مراجعه کنید.

بررسی نحوه تاخوردگی توالی ها و تشکیل ساختار های PDB

گروه دیگری از دانشمندان، درست برعکس موضوع بالا، در حال بررسی این موضوع هستند که چه ساختار های پروتئینی شناخته شده ای می توانند بهترین مدل برای بررسی نحوه تاخوردن یک توالی خاص باشند. به این کار در اصطلاح threading می گویند. با گسترده شدن بانک های اطلاعاتی ساختار های پروتئینی، این روش روز به روز مفید تر خواهد بود، و مدل سازی ها به جای اینکه از روی ساختار های پروتئینی مشابه صورت گیرد، از روی توالی هایی که زیاد هم به هم شبیه نیستند، انجام خواهد شد.

برای آشنایی بیشتر با این روش سایت های زیر را ببینید:

- ✓ **FUGUE:** www-cryst.bioc.cam.ac.uk/~fugue/
- ✓ **U CLA Fold recognition:** www.doe-mbi.ucla.edu/Services/FOLD/
- ✓ **The PROSPECT server:** compbio.ornl.gov/structure/prospect/

دینامیک و حرکت ساختار های پروتئینی

بعضی دانشمندان در تلاشند تا حرکاتی را که در ساختمان نهایی و پایدار پروتئین ها اتفاق می افتد را شبیه سازی کنند. دانستن دینامیک مولکول های پروتئینی، به درک جزئیات بر هم کنش های بین پروتئین ها و واکنش های آنزیم کمک می کند.

برای آشنایی بیشتر با دینامیک و حرکات پروتئین ها، سایت های زیر را ببینید:

- ✓ **The Brooks Lab site:** brooks.Scripps.edu/
- ✓ **The El Nemo site:** www.igs.cnrs-mrs.fr/elnemo/
- ✓ **The Database of Macromolecular Movements site:** molmovdb.mbb.yale.edu/MolMovDB/

پیش گویی برهم کنش ها

در نهایت مشکل دیگری که دانشمندان برای حل آن تلاش می کنند، اینست که آیا دو مولکول پروتئینی یا یک پروتئین و یک مولکول کوچک دیگر، ممکن است برهم کنش داشته باشند یا خیر؟ و در صورتی که جواب مثبت باشد، این برهم کنش به چه شکلی صورت می گیرد؟ به این زمینه در اصطلاح molecular docking می گویند. در اغلب موارد ساختار پروتئینی دو پروتئین پس از برهم کنش، از روی ساختار تک تک آن ها با استفاده از کامپیوتر قابل پیش بینی نیست.

برای آشنایی بیشتر به سایت های زیر مراجعه کنید:

- ✓ **AutoDock:** www.scripps.edu/mb/olson/doc/autodock/
- ✓ **FlexX:** www.biosolveit.de/FlexX/
- ✓ **FTDock:** www.bmm.icnet.uk/docking/
- ✓ **Hex:** www.csd.abdn.ac.uk/hex/
- ✓ **HotDock:** wwwcs.uni-paderborn.de/~lst//HotDock/index.html

فصل ۱۲

کار کردن با RNA

در این فصل می خوانیم:

- ✓ نحوه جستجوی بانک های اطلاعاتی مولکول های RNA
- ✓ پیش گویی ساختار دوم RNA
- ✓ یافتن مولکول های RNA غیرکدکننده

.....

"زیست شناسی حداقل به اندازه ۵۰ سال دیگر مطالب جالب توجه دارد."

— James D. Watson

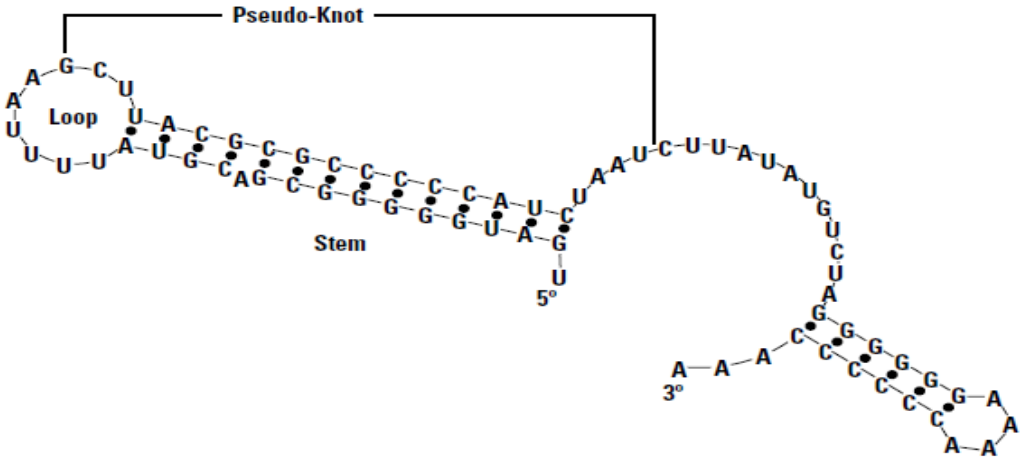
در این فصل یاد می گیرید که چگونه با استفاده از کامپیوتر سحرانگیز ترین مولکول یعنی RNA را آنالیز کنید.RNA یک ملکول شگفت انگیز است چرا که هر آنچه که شما فکرش را بکنید دارد: اطلاعات ژنتیکی (درست مثلDNA) و نقش کاتالیزگری (درست مثل پروتئین ها). در واقع هر چه دانشمندان بیشترRNA را بررسی میکنند بیشتر تعجب میکنند که چرا طبیعت زحمت ساخت پروتئین ها و DNA را به خود داده است! از آنجائیکه RNA بسیار قابل پیش بینی تر از DNA و پروتئین است آنالیز آن با کامپیوتر آسان تر است و اطلاعات بیشتری میتوان از آن بدست آورد.

در بخش اول این فصل ما به شما نشان خواهیم داد که چگونه از mfold (یک سرور مشهور) برای پیش گوئی و نمایش ساختمان دوم یک توالی RNA استفاده کنید. همچنین به شما نشان میدهیم که چگونه اطمینان حاصل کنید که به ساختار درست دست پیدا کرده اید.

در بخش دوم فصل روشهای جستجوی اختصاصی RNA ها در بانک های اطلاعاتی را خواهید آموخت . اگر ساختمان دوم RNA مورد نظر خود را بدانید، می توانید در بانک های اطلاعاتی به دنبال ساختارهای مشابه بگردید، حتی ساختارهایی که شباهت آنها با RNA مورد نظر شما بسیار کم باشد. این روش یک روش بسیار قدرتمند است برای کشف اعضای جدید یک خانواده ژنی یا اعضای جدید یک خانواده از RNA که خود شما کشف کرده اید. بخش پایانی این فصل توضیح کوتاهی است در مورد مولکولهای کوچک RNA به نام siRNA یا miRNA معروفند. همچنین منابع اطلاعاتی مرتبط با این زمینه و نحوه استفاده کارآمد از آنها را معرفی خواهیم کرد. متأسفانه سایت های بزرگ و قدرتمند مثل NCBI یا EBI به RNA به اندازه زمینه های دیگر اهمیت نداده اند؛ اما خوشبختانه انجمن RNA در این زمینه بسیار فعال است. در اینترنت منابع با کیفیت زیادی وجود دارد که در دسترس عموم قرار دارند. در بخش انتهایی این فصل بعضی از این منابع معرفی خواهند شد.

پیش گویی، مدل سازی و طراحی ساختارهای دوم RNA

یک پیشرفت بزرگ در دهه ۱۹۷۰ کشف این واقعیت بزرگ بود که RNA می تواند ساختارهای دوم و سوم پیچیده داشته باشد (شکل ۱-۱۲). مطلب جالب توجه در مورد این ساختارها اینست که قوانین حاکم بر آنها بسیار ساده تر و قابل پیش بینی تر از پروتئین هاست. اساس و پایه بسیاری از ساختارهای دوم RNA قانون جفت شدن بازهای آلی واتسون و کریک است.



شکل ۱-۱۲: شکل معمول ساختار دوم RNA

ساختار دوم RNA

یک مولکول RNA تک رشته ای درست مانند یک نوار چسب عمل می کند. این مولکول تا زمانیکه بعضی از بازهای آلی آن در برابر آب محافظت نشوند، پایدار نخواهد بود. یک راه برای محافظت یک باز آلی RNA از حمله مولکولهای حلال اینست که با یک باز آلی دیگر جفت شود. البته تشکیل همه جفت بازهای ممکن به یک میزان مفید نخواهد بود. به عنوان مثال جفت شدن گوانین با سیتوزین پایدار کننده تر از جفت شدن آدنین با اوراسیل است.

با جفت شدن این بازها ساختار دوم RNA به وجود می آید. اگر یک مولکول حاوی دو قسمت بلند مکمل همدیگر باشد، یک ساقه[1] بسیار پایدار به وجود می آید (شکل ۱-۱۲). محل های جفت نشده بین دو ساقه تشکیل حلقه[2] می دهند. همیشه لازم نیست که دو رشته تشکیل دهنده یک ساقه کاملا مکمل یکدیگر باشند بلکه ممکن است حاوی بازهای جفت نشده نیز باشند که این قسمت ها را جوانه[3] می نامند.

فرض می کنیم که مولکول RNA بطور طبیعی تمایل دارد با تشکیل مجموعه ای از جفت بازها به پایدارترین حالت خود برسد. این مفهوم با نام "مدل کمترین سطح انرژی[4] " نیز خوانده می شود. چنین ساختاری از RNA سطح انرژی منفی دارد (مثلا Kcal/mol 70-)، بنابراین برای به هم زدن ساختمان آن نیاز به انرژی (گرما) است. چگونگی رسیدن مولکول های RNA و همچنین پروتئین ها به کمترین سطح انرژی هنوز دقیقا مشخص نیست، اما مشخص شده است که این فرایند خیلی سریع درون سلول اتفاق می افتد. میزان پایداری یک ساختار دوم پروتئین فقط وابسته به محتوای جفت بازهای GC نیست.

[1] stem
[2] loop
[3] bulge
[4] the Lowest Energy Model

عوامل زیادی مانند روی هم انباشته شدن جفت بازها (که باعث ساخته شدن ساقه های بسیار پایدارتر از بقیه می گردد)، و اندازه حلقه ها نیز مؤثرند. الگوریتم های پیچیده و دقیق (مثل الگوریتم های برنامه mfold) می توانند این عوامل را در نظر گرفته و یک ساختار بهینه را مدل سازی کنند.

برهم کنش های دیگری که منجر به ایجاد ساختمان سوم می شوند نیز در پایداری مولکول نقش دارند. این ساختمان ها شامل شبه گره هایی[1] هستند که از برهم کنش با محدوده بلند بین یک حلقه و یک بخش دیگر از مولکول RNA به وجود می آیند (شکل ۱-۱۲). برهم کنش بین مولکول های RNA و دیگر عناصر شیمیایی مثل یونها، پروتئین ها و دیگر مولکول های RNA نیز نقش مهمی در پایداری مولکول RNA دارد.

متأسفانه پیش گویی برهم کنش های نوع سوم یا اثر پروتئین ها بر نحوه تا خوردن مولکول RNA مشکل است. بنابرین هنگامی که شما ساختمان سوم RNA خود را پیشگویی می نمایید، این پیشگویی معمولا به این فرض بستگی دارد که درون سلول این RNA روی خودش پیچ می خورد. چون این فرض هیچ وقت صادق نیست، همیشه این احتمال وجود دارد که پیشگویی شما نسبتا یا کاملا غلط باشد. قاعده کلی این است که پایدارترین ویژگیها از نظر پایداری منطقا محتمل تر می باشند.

استفاده از برنامه mfold

mfold با استفاده از الگوریتم Zuker می تواند بهینه ترین ساختار دوم یک مولکول RNA را از نظر سطح انرژی پیش گویی کند. این برنامه از یک مدل پیچیده استفاده می کند که قادر است بسیاری از عوامل فیزیکی اثر گذار بر نحوه تا خوردن مولکول RNA را مد نظر قرار می دهد (مثلا pH، دما و سوگیری ناشی از ترکیب بازهای RNA در یک ناحیه مورد نظر).

به یاد داشته باشید که مدل انرژی که mfold از آن استفاده می کند، بر هم کنش های شکل سه بعدی مولکول یا بر هم کنش پروتئین-RNA را، که می توانند یک ساختار نیمه پایدار را پایدار کنند، در نظر نمی گیرد. پس همواره این احتمال وجود دارد که پیش گویی mfold با آنچه در واقعیت وجود دارد مطابق نباشد.

برای اینکه شما بهتر بتوانید در مورد پایداری مولکول تصمیم بگیرید، mfold ساختارهای نیمه پایدار را بر می گرداند تا خودتان در مورد پایداری آن ها قضاوت کنید. بعلاوه شما می توانید به mfold دستور بدهید که یک بر هم کنش خاص را، که مطمئن هستید اتفاق می افتد، مد نظر قرار دهد.

همچنین توجه کنید که برنامه mfold شبه گره ها را نیز نمی تواند مد نظر قرار دهد.

برای کار کردن با mfold در اولین گام باید یک توالی RNA داشته باشید. ممکن است بپرسید: "از کجا؟" . اگر دنبال RNA های غیرکدکننده[2] می گردید، بانک های اطلاعاتی DNA منبع خوبی برای این کار نیستند. بانک های اطلاعاتی اختصاصی RNA هم کم هستند و این بانک ها هم اغلب اختصاص به RNA ریبوزومی دارند. در بخش انتهایی این فصل تعدادی از این سایت ها معرفی می شوند (جدول ۲-۱۲).

بهترین اندازه برای توالی RNA چقدر است؟ سؤال خوبی است؛ با افزایش طول توالی، زمان لازم برای آنالیز آن توسط mfold به سرعت افزایش می یابد. به عنوان مثال اگر آنالیز یک توالی به طول ۱۰۰ باز یک ثانیه طول بکشد، آنالیز یک توالی ۲۰۰ بازی ۱۰ ثانیه طول خواهد کشید. در حال حاضر حد بالایی طول توالی برای آنالیز در mfold، ۳۰۰۰ باز است که برای RNA های غیرکدکننده مناسب است و فقط RNA های ریبوزومی خیلی بزرگ در این محدوده جای نمی گیرند.

[1] pseudo-knots
[2] non-coding RNA

برای استفاده از برنامه mfold مراحل زیر را دنبال کنید:

۱- وارد آدرس اینترنتی lowelab.ucsc.edu/GtRNAdb/Haem_infl/Haem_infl-tRNAs.fa زیر شوید. لیستی از توالی های tRNA مربوط به ژنوم هموفیلوس آنفلوانزا که توسط Sean Eddy's tRNA scan server پیش گویی شده است را خواهید دید. ترتیب توالی ها بر اساس قابلیت اطمینان پیشگویی آنهاست. بنابراین توالی هایی که در بالای لیست قرار گرفته اند تقریبا ۱۰۰ درصد مطمئن هستند.

۲- اولین توالی RNA را انتخاب کرده و آنرا کپی کنید (شکل ۲-۱۲).

۳- وارد آدرس اینترنتی www.bioinfo.rpi.edu/applications/mfold/ شوید.

صفحه Applications از سایت Center for Bioinformatics (Rensselaer Polytechnic Institute) ظاهر می شود.

شکل ۲-۱۲: انتخاب کردن یک توالی tRNA

شما همچنین می توانید از دو نسخه اروپایی mfold نیز استفاده کنید:

bibiserv.techfak.uni-bielefeld.de/cgi-bin/mfold_submit

bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html

یا نسخه استرالیایی آن:

mfold.burnet.edu.au

۴- در پایین صفحه روی لینک RNA Folding کلیک کنید.
صفحه اصلی سرور mfold ظاهر می شود (شکل ۳-۱۲).

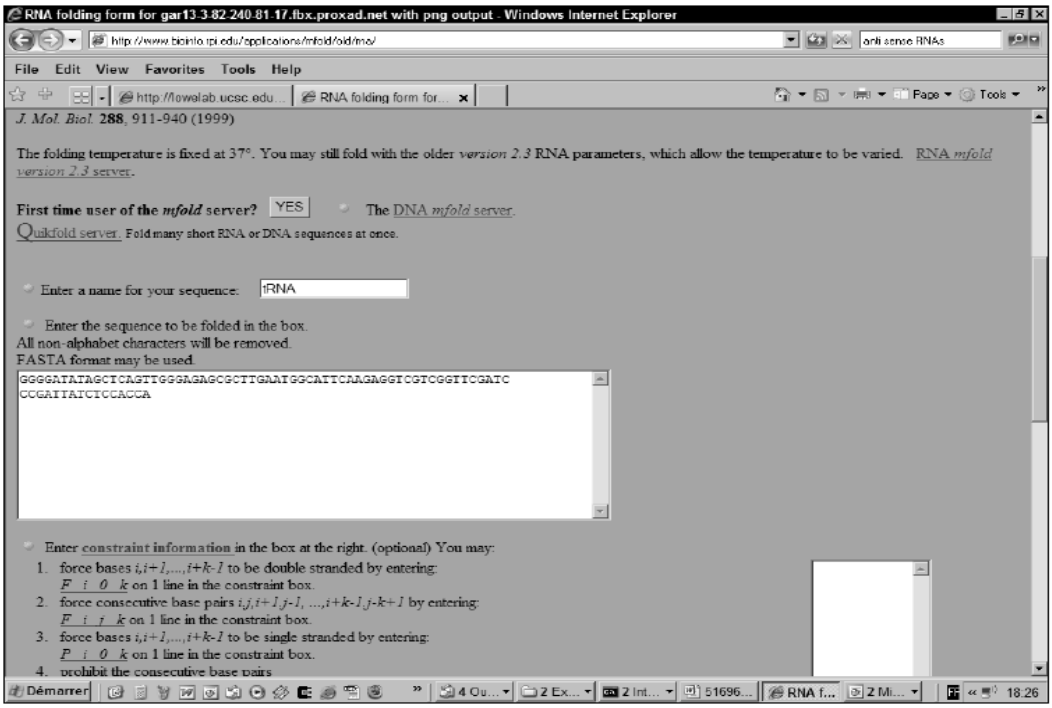
۵- یک نام برای توالی خود در قسمت مربوطه وارد کنید.

مثلا ما توالی خود را tRNA نامیدیم.

۶- در پنجره Sequence نام توالی خود را که قبلا کپی کرده بودید، وارد کنید.

بقیه پارامترها را می توانید بدون تغییر باقی بگذارید.

در ادامه این فصل در قسمت "وارد کردن اثر برهم کنش های RNA در mfold" یاد می گیرید که چگونه محدودیت های تعریف شده ای را در برنامه mfold اعمال کنید.



شکل ۳-۱۲: صفحه اصلی سرور mfold

۷- کمی پایینتر در صفحه روی گزینه Fold RNA کلیک کنید.

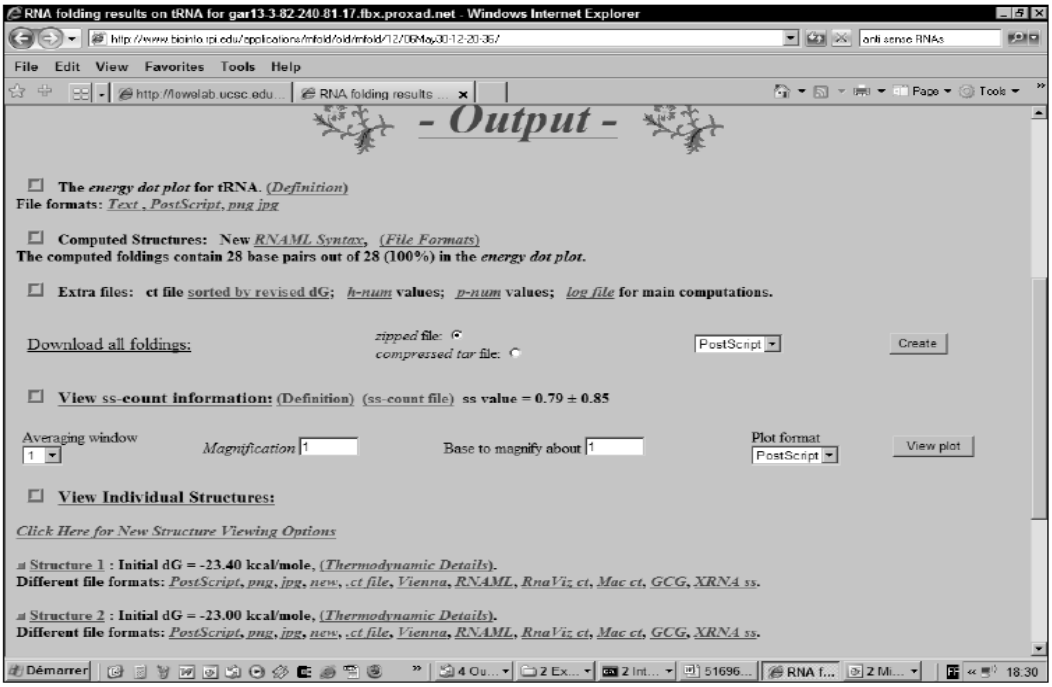
صفحه Output ظاهر می شود (شکل ۴-۱۲). این صفحه بسیار پیچیده تر بنظر می رسد، چرا که حاوی لینک های زیادی است که هر کدام نتایج مورد نظر شما را به یک فرمت خاص نشان می دهند. همچنین در این قسمت اطلاعات ترمودینامیک مربوط به مدل شما نیز وجود دارد که با استفاده از آنها می توانید پایداری مدل را محاسبه کنید.

۸- تصویر jpg بهترین مدل را در کامپیوتر به نمایش بگذارید.

أ- در صفحه Output پایین بیایید تا به بخشی با عنوان View Individual Structures برسید (شکل ۴-۱۲). در این قسمت ساختارها بر اساس میزان پایداری مرتب شده اند. به عنوان مثال ساختار شماره ۱ بیشترین پایداری را دارد، ساختار شماره ۲ از نظر پایداری در رتبه دوم قرار دارد و ...

ب- در قسمت ساختار ۱ لینک های مربوط به ساختارهای مختلف را پیدا کرده و روی لینک jpeg کلیک کنید. صفحه جدیدی باز می شود که پایدارترین ساختار دوم مربوط به توالی شما را نشان می دهد (شکل ۵-۱۲). البته ممکن است این ساختار پایدارترین نباشد اما یک کاندیدای خوب برای پایدارترین ساختار است.

۹- با راست کلیک روی شکل و انتخاب گزینه Save as این شکل را ذخیره کنید.



شکل ۴-۱۲: صفحه output سایت mfold

تفسیر نتایج mfold

زمانیکه یک توالی به mfold داده می شود، این برنامه ساختارهای ثانویه با کمترین سطح انرژی را جستجو می کند. البته صحت ساختار ثانویه به دست آمده به صحت مدل انرژی بستگی دارد؛ اگر این مدل صحیح نباشد، ساختار به دست آمده نیز صحیح نخواهد بود.

قانون کلی در برنامه mfold اینست که هر چه ساختار پایدارتری پیش گویی شود، احتمال اینکه برهم کنشی که منجر به ایجاد این ساختار می شود واقعی باشد، بیشتر است. در صفحه Output همچنین برای درک بیشتر ماهیت پایداری مولکول RNA چندین نتیجه مرتبط با پایداری توالی مورد نظر شما آورده شده است. در بین این نتایج بهتر است به مواردی که در زیر آمده است، بیشتر توجه کنید:

- ✓ The Energy Dot Plot: این قسمت ساقه هایی که بخشی از ساختار دوم توالی مورد نظر شما هستند را نشان می دهد.
  - ساقه ها به صورت یک خط سیاه عمود بر خط اصلی نشان داده شده اند.
  - برهم کنش بین نوکلئوتیدهایی که در توالی فاصله زیادی با هم دارند[1]، به صورت خط هایی نزدیک به گوشه بالای راست یا پایین چپ نشان داده شده اند.
  - خطوط رنگی نشاندهنده ساختارهای ثانویه کمتر بهینه هستند. تغییر رنگ از قرمز به زرد به ترتیب بهترین و بدترین ساختارها را نشان می دهد.
- ✓ بخش View Individual Structures: این بخش حاوی اطلاعاتی در مورد ساختارهای مختلفی است که mfold محاسبه کرده است. بهترین ساختار در بالاترین قسمت قرار گرفته است. در این بخش همچنین انرژی آزاد

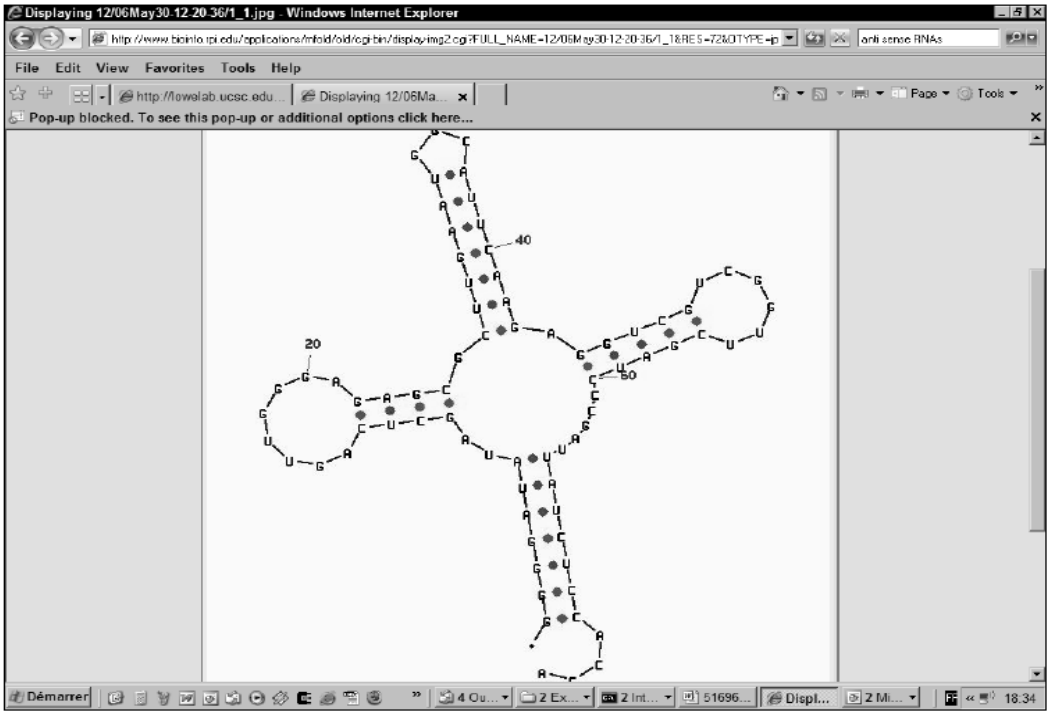
[1] Distal interactions

ساختار مورد نظر شما ذکر شده است. هر چه این انرژی کمتر باشد مطلوب تر است. مثلا در مورد یک مولکول به طول ۲۰۰ باز، انرژی آزاد باید کمتر از Kcal 50- باشد تا چنین مولکولی قابل پذیرش باشد.

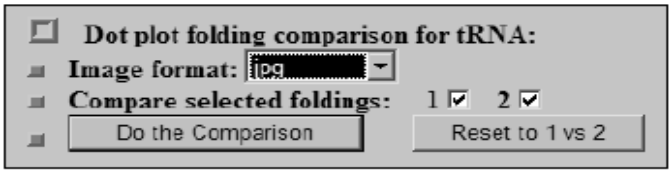
- ✓ بخش Dot Plot Folding Comparisons: در این بخش بسیار جالب، شما می توانید چند ساختار کمتر بهینه را با هم مقایسه کنید. بدین منظور مراحل زیر را دنبال کنید:
  ۱. با کلیک راست بر روی Image Format یک منو باز می شود؛ گزینه jpg. را انتخاب کنید.
  ۲. در قسمت Compare Selected Foldingw ساختارهایی را که می خواهید مقایسه کنید انتخاب کنید.
  ۳. روی دکمه Do the Comparison کلیک کنید (شکل ۶-۱۲).

  شکلی شبیه به شکل ۵-۱۲ ظاهر می شود. این شکل عناصری از ساختارهای ثانویه را که بین ساختارهای انتخاب شده مشترک هستند را نشان می دهد. عناصر رنگی مربوط به ساقه هایی هستند که فقط در ساختارهای دوم کمتر بهینه دیده می شوند.

در صورتیکه ساختار بهینه و ساختارهای کمتر بهینه سازگاری بالایی با هم داشته باشند، احتمال اینکه ساختار بهینه همان ساختار بیولوژیک واقعی باشد افزایش می یابد.



شکل ۵-۱۲: نمایش یک پیش گویی ساختار دوم RNA



شکل ۶-۱۲: مقایسه چند ساختار ثانویه در mfold

پیش گویی ساختار دوم یک توالی به تنهایی نتایج جالبی میدهد. اما زمانی پیشگویی از روی تطابق چند توالی صورت گیرد بهترین نتایج بدست خواهد آمد. این پدیده را پدیده covariance می گویند.

متأسفانه پیشگویی های مبتنی بر تطابق چند توالی در اینترنت بصورت رایگان در دسترس عموم نیست. اما اگر یک کامپیوتر قدرتمند دارید و میتوانید چند نرم افزار را روی آن نصب کنید، میتوانید از نرم افزاری به نام cove استفاده کنید که در آدرس اینترنتی زیر در دسترس است:

www.genetics.wustl.edu/eddy/software/#cove

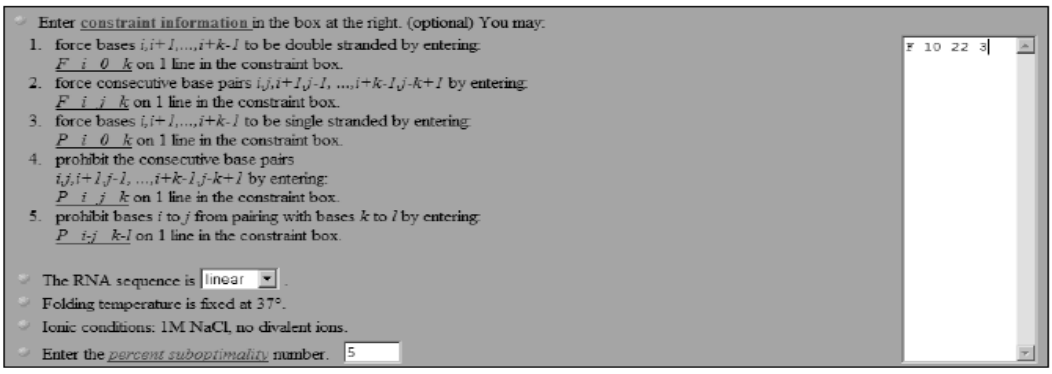
همچنین چند روش در سایت کتابخانه Robin Gutell به آدرس اینترنتی www.rna.icmb.utexas.edu وجود دارد. میتوانید در این سایت ثبت نام و از امکانات آن استفاده کنید.

وارد کردن اثر بر هم کنش های RNA در mfold

اگر بر اثر تجربه میدانید که چه بازهایی در توالی RNA مورد نظر شما با هم جفت میشوند، میتوانید این اطلاعات را به mfold بدهید تا از آن ها استفاده کند. مثلا می توانید به mfold بگویید که نوکلئوتیدهای ۱۰ و ۱۱و ۱۲ به ترتیب با نوکلئوتیدهای ۲۲ و ۲۱ و ۲۰ جفت میشوند، برای این کار در فرم بر هم کنش در mfold بنویسید:

F, 10, 22, 3

این دستورالعمل یعنی شما به mfold دستور داده اید (F=Force) که یک برهم کنش را انجام دهد که با جفت شدن نوکلئوتید ۱۰ با ۲۲ شروع می شود و ساقه ای به طول ۳ نوکلئوتید را بوجود می آورد. شکل ۷-۱۲ شکل صحیح پر کردن فرم بر هم کنش در mfold را نشان میدهد.



شکل ۷-۱۲: وارد کردن بر هم کنش های ثانویه در mfold

جستجوی بانک های اطلاعاتی و ژنوم ها برای یافتن توالی های RNA

در یک سلول همه ژنها به مولکول های RNA نسخه برداری می شوند و بدین ترتیب سلول از ژنها استفاده می کند. نکته ای که گاهی مورد توجه قرار نمی گیرد اینست که دو نوع RNA وجود دارد : انواعی که به پروتئین ترجمه می شوند و انواعی که مستقیما توسط سلول مورد استفاده قرار میگیرند که RNA های غیر کدکننده نیز نامیده می شوند.

RNA های غیر کد کننده یک خانواده بزرگ را تشکیل میدهند و هر ساله تعدادی عضو جدید هم به آنها اضافه می شود. هر چه مکانیسم پایه عملکردهای سلول بیشتر شناخته می شود، وسعت نقش RNA های غیر کد کننده در فعالیت های سلول نیز بیشتر روشن می گردد. این مولکول (در کارهایی مثل نسخه برداری، Splicing، همانندسازی و احتمالا بسیاری از کارهایی که هنوز ما از آنها اطلاع نداریم نقش دارند. بنظر میرسد که همه مکانیسم های پایه حیات به این مولکول های کوچک وابسته است و گفته میشود که بخش بزرگی از زیست شناسی آینده به آنها اختصاص خواهد داشت.

متأسفانه بجز RNA ریبوزومی سایر RNA های غیر کد کننده معمولا کوتاه و حفاظت نشده هستند. معمول ترین ویژگی آنها ساختار دوم آنهاست. بنابر این یافتن آنها با روش های تطابق و همولوژی کار مشکلی است (در فصل ۷ اطلاعات بیشتری در مورد این روش ها ارائه شده است).

روش هایی که در این فصل به شما نشان میدهیم نه تنها با استفاده از توالی RNA بلکه با استفاده از اطلاعات مربوط به ساختارهای ثانویه شناخته شده، بانک های اطلاعاتی را جستجو می کند. اگر با استفاده از mfold چگونگی تا خوردن یک خانواده RNA را پیش گویی کرده اید، می توانید با استفاده از این اطلاعاتی که بدست آورده اید، اعضای جدید خانواده مذکور را نیز جستجو کنید .

در این بخش ابتدا به شما یاد می دهیم که چگونه با استفاده از tRNAScan تنها به دنبال مولکول های tRNA بگردید. همچنین متد PatScan را به شما معرفی می کنیم که با آن می توانید مولکول های RNA با ساختارهای ثانویه ای که خودتان اختصاصا تعیین کرده اید را جستجو کنید.

یافتن مولکول های tRNA در یک ژنوم

مولکول های tRNA انواع کوچک و غیر کد کننده RNA هستند که سلول برای ساخت پروتئین از آنها استفاده می کند. این مولکول ها بهترین مثال از ژنهای غیر کدکننده مهم هستند که یافتن محل آنها در ژنوم مشکل است. همه ژنوم های باکتریایی و یوکاریوتی حاوی tRNA هستند، اما متأسفانه شناسایی این ژنها با وجود فراوانی آنها در ژنوم ، مانند سایر RNA های غیر کد کننده کوچک مشکل است.

برای این کار تکنیک های بسیار پیچیده تر از BLAST جهت جستجوی بانک های اطلاعاتی لازم است. اما خوشبختانه تعدادی برنامه خیلی کارا برای پیش گویی ژنهای tRNA در ژنوم های یوکاریوتی و پروکاریوتی وجود دارد.

متد RNAscan-SE در سایت دانشگاه واشنگتن واقع در سنت لوئیز به آدرس Selab.janelia.org/tRNAscan-SE در دسترس است.

جستجوی الگوهای RNA با استفاده از PatScan

برنامه tRNAscan تنها زمانیکه در جستجوی tRNA ها باشید برنامه کاملی است: اما زمانیکه در جستجوی یک خانواده خاص از RNA باشید نیاز به یک ابزار جامع تر دارید که این امکان را به شما بدهد که از اطلاعات شخصی خود در مورد آن خانواده RNA، در جستجوی خود استفاده کنید. PatScan برای این منظور عالی است.

PatScan امکان جستجوی بانک های اطلاعاتی را با الگوهایی که به درستی می توانند ساختارهای ثانویه RNA را توصیف کنند، به شما می دهد. از دیدگاه کامپیوتر ای، این نوع جستجو بسیار گران است، به همین دلیل است که PatScanآدرس پست الکترونیک شما را می خواهد و نتایج جستجو پس از چند ساعت به دست شما می رسد.

اگر PatScan برنامه مطلوب شما نبود (مثلا بخاطر اینکه نمی خواهید توالی های شما وارد اینترنت شوند) میتوانید از RNAmot که شبیه PatScan است استفاده کنید. برای استفاده از این برنامه نیاز است که این نرم افزار را روی کامپیوتر خود نصب کنید، اما توجه داشته باشید که این نرم افزار برای افراد تازه کار مناسب نیست. این نرم افزار در آدرس اینترنتی زیر برای دانلود در

دسترس است:

www.esil.univ-mrs.fr/~dgaut/download/

هنگام استفاده از PatScan طراحی یک الگوی PatScan مهمترین مرحله است و بیشترین مهارت وفکر کردن را نیاز دارد. اگر می خواهید قوانین حاکم بر الگوهای PatScan را دقیقاً بدانید، می توانید اطلاعات آنلاین در این مورد را در آدرس اینترنتی زیر بخوانید:

www.unix.mcs.anl.gov/compbio/PatScan/HTML/matching_nucleotides.html

برای شروع طراحی الگوی مورد نظر خود ابتدا زمینه علاقمندی خود را مشخص کنید. لیست زیر در این زمینه به شما کمک می کند:

- ✓ از روی ساختارهای اولیه، عناصر ساختار ثانویه را تعیین کنید.
- ✓ ساقه ها را مشخص کرده وآنها را نامگذاری کنید. نام قسمت سمت چپ اولین ساقه p1 و قسمت سمت راست آن p1 ~ است.
- ✓ طول هر یک از عناصر را مشخص کنید.
- ✓ در هر یک از عناصر ساقه، جفت بازهای غیر متعارف واتسون و کریک را مشخص کنید.
- ✓ این ساختار را با زبان خودتان بازنویسی کنید، مثلا:

[stem p1 of length to 9] [Loop of length 3 to] [stem ~p1 of length to 9, c omplementary to the left stem p1]

- ✓ ساختار ثانویه را به الگوی PatScan تبدیل کنید.
- ✓ 9... = p1 نشاندهنده ساقه p1 با ۸ تا ۹ نوکلئوتید است.
- ✓ ۸..۳ نشاندهنده یک الگو با ۳ تا ۸ نوکلئوتید است.
- ✓ p1~ نشاندهنده ساقه ای است که به صورت عکس مکمل p1 است.

بنابراین می توانید الگوی خود را به صورت زیر بنویسید:

p1= ...9 3... ~p1

اگر می خواهید مشخصات چند ساقه را بنویسید به آنها نام های متفاوتی بدهید، مثلا p1، p2، p3، ...

ممکن است نیاز باشد ساقه های تودرتو نیز تعریف کنید، مثل:

p1= ...9 3... p2=5...7 4...5 ~p2 3... ~p1

- ✓ شما هر چیزی را که برای توصیف یک ساختار پیچیده RNA نیاز باشد می توانید در PatScan جای دهید؛ مثلا توانایی توصیف شبه گره ها و برهم کنش های با دامنه بلند بین یک حلقه و بخش دیگری از مولکول RNA که در شکل ۱-۱۲ نشان داده شده اند.

با دستورالعمل زیر کار کردن با PatScan را امتحان کنید:

۱. الگوی PatScan خود را آماده کنید.

۲. وارد آدرس اینترنتی www.unix.mcs.anl.gov/compbio/PatScan/HTML/scanner.html شوید.
صفحه اصلی PatScan ظاهر می شود.

۳. بانک اطلاعاتی را که می خواهید در آن جستجو کنید از قسمت مربوطه انتخاب کنید.
به عنوان مثال ما DNA:Human را انتخاب می کنیم.

Table 12-1 RNA-Specific Pattern Rules in PatScan

Rule	*Example*	*Description*
Pairing rule	r1={au,gc} p1=10...12 3...8 r1~p1	Indicates that the p1 pairing only involves au and gc pairs.
Mismatches	p1=10...12 3...8 r1~p1[1,4,2]	The pairing between the two p1 stems can involve at most: 1 mismatch in p1 Vs ~p1 4 unpaired symbols in p1 2 unpaired symbols in ~p1.
Specify patterns	p1=10...12 AUGC~p1	Specifies that the loop MUST contain only AUGC. Note that this pattern must be in uppercase, as opposed to the pairing rules.

جدول ۱-۱۲ تعدادی از قابلیت های برنامه قدرتمند PatScan را نشان می دهد.

در این سرور نوع بانک اطلاعاتی به صورت پیش نویش Swiss-Prot است، اما این بانک اطلاعاتی برای جستجوی ساختار ثانویه RNA مناسب نیست. شما باید یک بانک اطلاعاتی مربوط به DNA انتخاب کنید.

۴. الگوی خود را وارد پنجره Pattern کنید.
مثلا ما p1 ~ ...3 9... =p1 را انتخاب می کنیم.

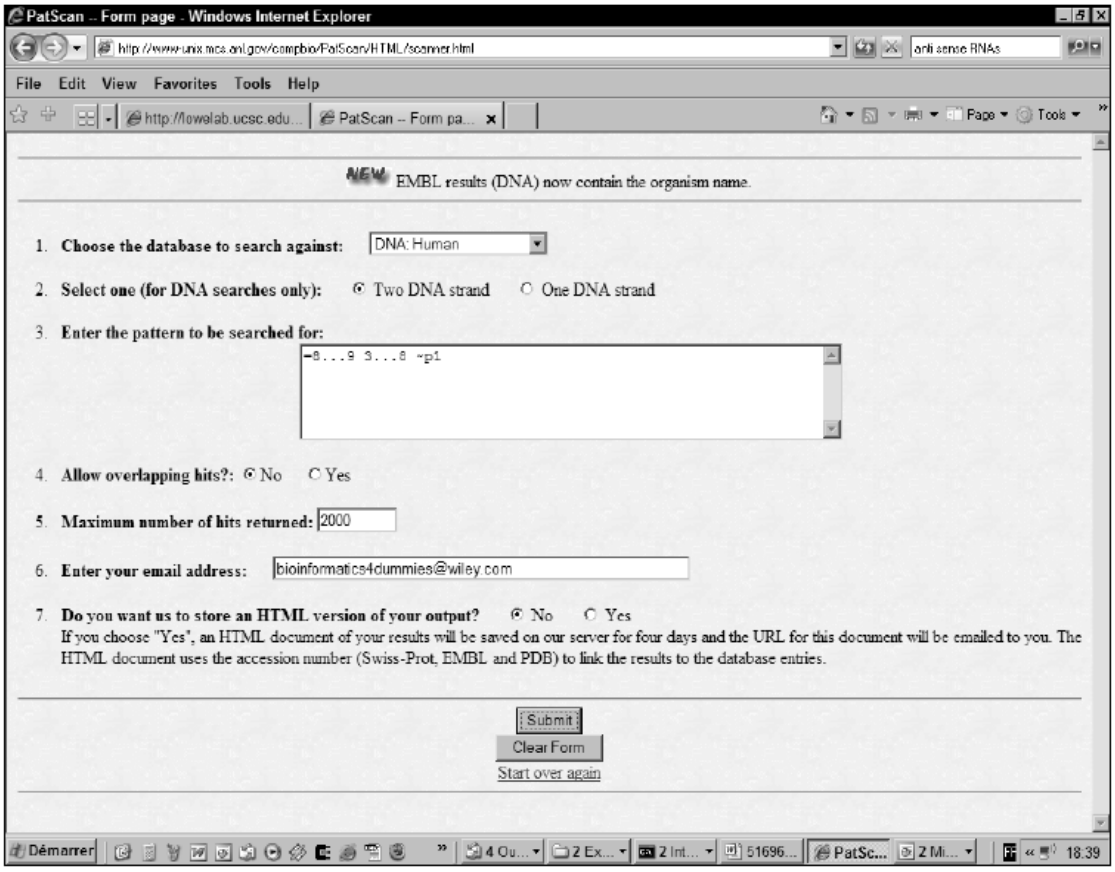
۵. آدرس پست الکترونیک خود را در قسمت مربوطه وارد کنید.

۶. روی گزینه Submit کلیک کنید.

بسته به اندازه بانک اطلاعاتی که انتخاب کرده اید و طول و میزان پیچیدگی الگوی شما، جستجو ممکن است تا یک ساعت طول بکشد.

۷. نتایج به دست آمده را آنالیز کنید (شکل ۹-۱۲).

در شکل ۹-۱۲ با استفاده از PatScan بخش Human از بانک اطلاعاتی GeneBank را جستجو کرده ایم. تمام الگوهای DNA که با الگوی ما تطابق دارند ظاهر می شوند و فراوانی آنها در GeneBank/EMBL/DDBL و همچنین مختصات آنها در توالی نیز ذکر می شوند.



شکل ۸-۱۲: صفحه اصلی سایت PatScan

اینکه یک توالی با الگوی شما مطابقت دارد، لزوماً به این معنی نیست که با ساختار ثانویه ای که در ذهن شماست، مطابق است! PatScan سازگاری اطلاعات موجود را با یک ساختار ثانویه خاص چک می کند، اما با این حال نمی توان با اطمینان گفت که ساختار به دست آمده پایدارترین ساختار برای این اطلاعات است.

```
Subject: PatScan Results

The results for your PatScan search:

DATABASE:
human

PATTERN:
p1=8...9 3...8 ~p1


Maximum # of hits requested:     2000
# of hits found/reported:        4

O60641:[530,549]  : AAAAAATT AATA AATTTTTT
O60641:[549,530]  : AAAAAATT TATT AATTTTTT
Q9NTY7:[525,544]  : AAAAAATT AATA AATTTTTT
Q9NTY7:[544,525]  : AAAAAATT TATT AATTTTTT
COMPLETED REQUEST
```

شکل ۹-۱۲: نتایج جستجوی PatScan

یافتن انواع جدید RNA: miRNA ها و siRNA ها

یکی از بزرگترین دستاوردهای زیست شناسی در سال های اخیر پدیده RNA interference یا RNAi در سلول های یوکاریوتی بوده است. این پدیده در واقع تداخل یک RNA غیرکدکننده ی کوچک و کوتاه با یک RNA کدکننده بزرگتر است که نتیجه آن ممانعت از بیان ژن مربوطه می باشد. این عمل با اتصال قطعه کوچک RNA به بخش مکمل زان در RNA بزرگتر اتفاق می افتد. به همین خاطر است که این RNA های کوچک گاهی antiscence RNA نیز نامیده می شوند. (برای اطلاعات بیشتر در زمینه نحوه جفت شدن بازهای مکمل به فصل ۱ و۲ مراجعه کنید.)

اگرچه امکان تداخل مولکول های RNA در محیط آزمایشگاهی، از قبل شناخته شده بود اما چیزی که همه را حیرت زده کرد این بود که این پدیده به طور طبیعی به وفور اتفاق می افتد.

در آن زمان RNA های طبیعی کوچک که توانایی غیرفعال کردن یک ژن را داشتند یکی پس از دیگری کشف می شدند. این RNA ها ابتدا در گیاهان یافت شدند و به خاطر توانایی آنها در غیر فعال کردن با به عبارتی خاموش کردن ژنها silencing RNAs (siRNAs) نامیده شدند. siRNA ها مختص گیاهان نیستند و در اکثر سلول های یوکاریوتی یافت شده اند.

زمانیکه دانشمندان در جستجوی siRNA ها بودند تعداد زیادی RNA شناخته نشده که هم اندازه siRNA ها بودند نیز کشف شدند. دانشمندان این RNA ها را microRNA یا miRNA نامیدند. هیچ کس نمی داند miRNAها در سلول چه نقشی دارند اما احتمال داده شده است که در تنظیم ژنها نقش داشته باشند. علاوه بر این، تعداد آنها در سلول نیز مشخص نیست چون یافتن آنها در ژنوم بسیاز سخت تر از یافتن پروتئین هاست. تنها خصوصیت مشترک miRNA ها کوچک بودن آنهاست. تعداد miRNA ها در ژنوم انسان نیز هنوز مشخص نیست.

بزرگترین وجه تمایز بین siRNA و miRNA اینست که siRNA دورشته ای اما miRNA تک رشته ای و معمولا دارای ساقه و حلقه است. این مولکول های RNA کوچک بخاطر توانایی شان در غیرفعال کردن ژنها به صورت کاملا اختصاصی، در بیولوژی مولکولی بسیار پرطرفدار شده اند و در صورتیکه راهی برای فرستادن کارا و دقیق آنها به محل های خاص در بدن یافت شود، از امیدبخش ترین نوآوری های پزشکی به شمار می روند.

جدول ۲-۱۲ اطلاعات کافی برای یافتن miRNA ها در یک ژنوم در اختیار شما قرار می دهد.

آنالیز رایگان RNA در اینترنت

اگر به آنالیز RNA علاقه دارید، همه لوازم مورد نیاز برای این کار در اینترنت وجود دارد. ادامه مطالب این فصل را بخوانید، تا درک بهتری از منابع آنلاین در دسترس در این زمینه پیدا کنید.

بررسی فرضیه تکامل با استفاده از RNA ریبوزومی

برای بررسی فرضیه تکامل نیاز است که ژنی را که در تمام موجودات زنده وجود دارد بشناسیم. سالهاست که دانشمندان RNA ریبوزومی (rRNA) را مسئول تکامل می دانند، طوریکه به محض اینکه یک باکتری جدید کشف می شود دانشمندان ابتدا rRNA آن را بررسی می کنند تا محل آنرا در درخت حیات پیدا کنند.

جدول ۳-۱۲ معروف ترین منابعی که مرتبط با rRNA هستند را لیست کرده است.

Table 12-2 Hunting Micro RNAs (miRNAs) over the Web	
Address	***Description***
sirna.cgb.ki.se/	An extensive collection of resources on silencing RNAs.
itb.biologie.hu-berlin.de/~nebulus/sirna/v2/	A database of all known human silencing RNAs.
microrna.sanger.ac.uk/sequences	The home of miRNAs at the Sanger Center in the UK. Probably one of the most extensive resources on micro-RNAs.
cbit.snu.ac.kr/~ProMiR2/	A resource for predicting miRNAs using probabilistic methods.
pictar.bio.nyu.edu	Prediction of the potential target of your miRNA on complete genomes.
bibiserv.techfak.uni-bielefeld.de/rnahybrid/	A resource for predicting the potential target of your miRNA on a user-provided genomic sequence.
mirna.imbb.forth.gr/microinspector/	Runs your genomic sequence against an exhaustive database of miRNAs.

یافتن RNA های غیرکدکننده ی کوچک

RNA های غیرکدکننده کوچک در بانک های اطلاعاتی مرجع و اصلی زیاد مورد توجه قرار نگرفته اند، اما خوشبختانه بانک های اطلاعاتی فرعی خوبی در این زمینه وجود دارد.

جدول ۴-۱۲ تعدادی از این منابع را نشان می دهد.

Table 12-3 Ribosomal RNA Resources on the Internet	
Address	***Description***
rdp.cme.msu.edu	Provides data and services, including the possibility to make online phylogenetic analysis.
www.psb.ugent.be/rRNA/lsu/	A European database on the larger of the two ribosomal subunits. It contains predicted structures. It is possible to query the database online. Features lots of online software.
www.psb.ugent.be/rRNA/ssu/	The "other" European database, this time dedicated to the small ribosomal subunit.

Table 12-4	Some Non-Coding RNA Resources
Address	***Description***
`condor.bcm.tmc.edu/smallRNA/smallrna.html`	Dedicated to small non-coding RNAs.
`rna.wustl.edu/tRNAdb/`	Dedicated to tRNAs.
`bighost.area.ba.cnr.it/BIG/UTRHome/`	Dedicated to the untranslated regions of genes.
`www.indiana.edu/~tmrna/`	Dedicated to the recently discovered *tmRNAs* that are both transfer and messenger RNAs. (If you don't yet know what this is, you MUST take a look at this fascinating Web site!)

منابع عمومی RNA

جدول ۵-۱۲ منابع عمومی RNA را لیست کرده است. این منابع بانک های اطلاعاتی، نرم افزارها و سرورهای آنلاین مرتبط با RNA را در اختیار شما قرار می دهد.

Table 12-5	A List of Generic RNA Resources
Address	***Description***
`bioinfo.lifl.fr/rna/`	A site dedicated to the detection of non-coding RNAs
`www.imb-jena.de/RNA.html`	RNA World, one of the most complete sites currently available
`www.rnabase.org/links/`	Another very complete list of sites

فصل ۱۳

طراحی درخت فیلوژنتیک[1]

در این فصل می خوانیم:

- ✓ کاربرد درخت های فیلوژنتیکی چیست؟
- ✓ انجام تطبیق چند توالی در یک درخت فیلوژنتیکی
- ✓ تخمین فاصله بین توالی ها
- ✓ ساختن یک درخت Neighbor Joining
- ✓ چگونه خودتان به تنهایی کیفیت یک درخت را ارزیابی کنید؟
- ✓ یافتن منابع مرتط با درخت های فیلوژنتیکی در اینترنت

.....

" بدون نظریه تکامل هیچ چیز در زیست شناسی مفهوم ندارد."

— Theodosius Dobzhansky (1973)

هدف این فصل این است که به شما نشان دهیم چگونه با استفاده از کامپیوتر و تعدادی از سرویس های اینترنتی یک آنالیز فیلوژنتیکی انجام دهید. این آنالیز به شما این امکان را می دهد که فرایند تکاملی یک گروه از موجودات زنده یا توالی ها را دوباره نویسی کنید. اگر یک دید کلی در مورد درخت های فیلوژنتیکی[1] دارید اما نمی دانید چگونه یک درخت بسازید، این فصل به شما کمک خواهد کرد.

در این فصل خواهید آموخت که چگونه یک توالی را انتخاب کنید، آنرا به فرم تطابق چند توالی[2] درآورید، و نهایتا چگونه آنرا به یک درخت (در واقع یک دیاگرام که شبیه درخت است) تبدیل کنید. سه روش برای این کار به شما ارائه خواهیم کرد:

- ✓ **ClustalW**

متد ساده ای است که باید به عنوان یک جعبه سیاه از آن استفاده کنید. در واقع شما قادر به کنترل این متد نیستید، اما می توانید مطمئن باشید که این برنامه از روش Neighbor Joining که روش نسبتا جاافتاده ای است استفاده می کند.

- ✓ **Phylip**

یک روش پیچیده تر که این امکان را به شما می دهد که پارامترهای اساسی در ساخت درخت را کنترل کنید.

[1] phylogenetic tree
[2] multiple sequence alignment

- ✓ **PhyML**

یک روش با صحت بسیار بالا که با حداکثر احتمال ممکن کار می کند و متخصصین این زمینه آنرا صحیح ترین روش می دانند. همانطور که از نام آن مشخص است، روش Maximum Likelihood سعی می کند درختی را که با بیشترین احتمال با تطابق توالی های شما مرتبط است، بسازد.

به خاطر داشته باشید که معمولا کیفیت تطابق چند توالی عامل محدود کننده در این روش است.(اگر در زمینه تطابق چند توالی اطلاعاتی ندارید به فصل ۹ مراجعه کنید.)

کاربردهای درخت فیلوژنتیکی

هدف مطالعات فیلوژنتیکی یا " فیلوژنی[1] " بازسازی تاریخچه حیات و توضیح گوناگونی موجودات زنده است. این روند و تاریخچه را می توان به صورت یک درخت بزرگ نشان داد (درخت حیات). اصل اساسی در فیلوژنی دسته بندی موجودات زنده بر مبنای میزان شباهت آنها به یکدیگر است، در واقع ما فرض می کنیم که هر چه دو گونه بیشتر به هم شبیه باشند (مثلا انسان و میمون)، جد مشترک آنها در درخت فیلوژنتیکی به آنها نزدیکتر است. فیلوژنی یک مبحث جدید نیست و خواه منشأ زیست شناسی نوین را نظریه داروین بدانیم یا دسته بندی Aristotle، نمی توانیم این حقیقت را نادیده بگیریم که زیست شناسی به میزان زیادی به دسته بندی موجودات وابسته است و بهترین راه دسته بندی استفاده از فیلوژنی است. فیلوژنتیک نوع خاصی از فیلوژنی است که مبنای آن مقایسه ژنهای هم پایه در گونه های مختلف و یافتن میزان ارتباط و نزدیکی گونه ها به هم است. حتی با روش های فیلوژنتیکی می توان ژنهای مختلف یک خانواده را بررسی و تاریخچه تکاملی ژنهای یک خانواده را نیز پیدا کرد. البته توجه داشته باشید که این درخت ها زمانی برای شما مفهوم پیدا می کنند که نظریه تکامل را قبول داشته باشید.

در سطح مولکولی نیز مطالب زیادی وجود دارند که با علم فیلوژنتیک قابل کشف هستند. وقتی تاریخچه یک خانواده از پروتئین ها را بررسی می کنیم، رد پای جهش های ژنی، حذف ها، دوبرابر شدن ها، پیدایش دو گونه از یک گونه، از دست دادن یا بدست آوردن یک عملکرد خاص، غیر فعال شدن و دیگر وقایعی که اثرات آنها باعث پیدایش جهان به شکل امروزی اش شده است. در آنالیزهای بیوانفورماتیک به سه دلیل عمده نیاز داریم از علم فیلوژنتیک استفاده کنیم:

- ✓ **تعیین نزدیکترین خویشاوندان یک ارگانیسم**

به عنوان مثال زمانیکه شما می خواهید یک باکتری جدید را بررسی کنید، می توانید توالی RNA ریبوزومی آن را تعیین کرده و آن را در یک درخت فیلوژنتیکی که حاوی همه انواع RNA های ریبوزومی شناخته شده است، قرار دهید. با این کار می توانید در مورد نوع این باکتری قضاوت خوبی داشته باشید.

- ✓ **کشف عملکرد و نقش یک ژن**

هنگام مطالعه یک ژن شما می توانید با استفاده از درخت های فیلوژنتیکی مطمئن شوید که آیا ژن مورد نظر شما با یک ژن شناخته شده دیگر از یک گونه دیگر اورتولوگ[2] است یا نه؟

- ✓ **بررسی دوباره منشأ یک ژن**

اکثر ژنهای یک خانواده روند تکامل را همراه با هم طی می کنند، اما گاهی ممکن است یک ژن در طی این روند از یک خانواده خارج و به یک خانواده دیگر وارد شود. درخت های فیلوژنتیکی برای یافتن چنین وقایعی بسیار کارآمد هستند. این وقایع را انتقال افقی[3] گویند.

---

[1] phylogeny
[2] Orthologous
[3] Horizontal transfer

دو ژن اورتولوگ ژنهایی هستند که متعلق به دو موجود زنده متفاوت اند اما از یک ژن مشترک منشا گرفته انـد و جـدایی دو گونه ارگانیسم باعث جدایی آن ها شده است (کاملا بر خلاف واقعه ای که "دو برابر شدن ژنی" نامیده می شود.). از نظر تئوری دو ژن اورتولوگ، در دو موجود زنده عملکرد های یکسانی دارند. از طرفی، زمانی کـه دو ژن طـی پدیـده دو برابر شدن ژنی بوجود آیند، آنها را پارالوگ [1] می نامیم. دو ژن پارالوگ متعلق به یک خانواده هستند، ولی نسبت به ژنهای اورتولوگ با احتمال بیشتری ممکن است عملکرد های متفاوتی داشته باشند.

برای درک بهتر مطلب، اگر سکان هدایت یک هواپیمای بوئینگ ۷۴۷ را یک ژن در نظر بگیریم، می توان آن را با فرمان اتومبیل اورتولوگ دانست. همچنین چرخ جلوی سمت چپ اتومبیل با چرخ جلوی سمت راست آن اورتولوگ است.

آماده کردن اطلاعات فیلوژنیک

زیست شناسان از چندین دهه قبل، با وجود تکنیک های مختلف، از فیلوژنی استفاده می کرده اند. با استفاده از فیلوژنتیک [2] (مطالعات فیلوژنی روی داده های ژنی) کارهای علمی در این زمینه کمی آسان تر شده اند. اما به هر حـال تفکـرات زیـادی پیرامون این زمینه علمی مطرح هستند و اختلاف عقیده های قدیمی هنوز در دنیای فیلوژنی وجود دارند. پس توجه داشته باشید که وقتی یک روش را انتخاب کردید، باید کاستی های آن را هم بپذیرید! در این فصل روش هایی مطرح می شوند که در سال های اخیر کمترین اختلاف نظر در مورد آنها وجود داشته است.

مطلبی که همه با آن موافق هستند اینست که کیفیت داده ها در این مطالعات خیلی مهم است. اگر شما داده های خیلی خوب و مناسب تحلیل های فیلوژنتیک داشته باشید، با استفاده از هر روش منطقی، یک درخت منطقی بـه دسـت خواهیـد آورد. در مقابل، اگر داده های شما خوب نباشند، هیچ متدی، هر چند بسیار خوب باشد، نمی تواند داده های شما را تحلیل کند.

فیلوژنتیک با استفاده از ژنها تاریخ تکاملی موجودات را بازسازی می کند. تفسیر داده های حاصل از این مطالعـات بـسیار ساده تر از تفسیر داده های حاصل از مطالعه مورفولوژی موجودات (شکل و ساختار موجودات) است که پـیش از ایـن زیـست شناسان به آنها اتکا می کردند. همچنین مقایسه توالی های ژنی بسیار عینی تر [3] از مقایسه خصوصیات ظاهری موجودات است.

در مطالعات فیلوژنتیکی داده خوب یعنی یک تطابق چند توالی با صحت بالا که حاوی توالی هایی باشد که به درستی انتخاب شده اند. در اولین قسمت از این بخش، یاد می گیرید که چگونه داده های مناسب را به دست آورید.

انتخاب توالی مناسب برای یک درخت

زمانی که یک درخت فیلوژنتیکی می سازید، فرض ابتدایی اینست که توالی های کـه مقایـسه مـی شـوند، جـد مـشترک دارند. اگر توالی های شما به اندازه کافی مشابه باشند، این فرضیه منطقی به نظر می رسد. با استفاده از درخت حاصل می توانید تاریخ تکاملی توالی های مورد استفاده را بازسازی کنید و میزان خویشاوندی آنها را بررسی کنید.

البته از این دیدگاه، تکامل همواره واگرا است، بنابراین اگر دو توالی مشابه باشند، احتمـال اینکـه ایـن تـشابه ناشـی از تکامـل همگرا باشد، رد می شود. در بسیاری از موارد، این قانون درست است. اما در سطح مولکولی، تکامل همگرا هم وجود دارد؛ اما بیشتر یک استثناست تا یک قانون.

---

[1] Paralogous
[2] Phylogenetics
[3] objective

استفاده از توالی های DNA یا پروتئین

برای اثبات ارتباط بین دو توالی، لازم است زمانی که انشعاب ها از یک جد جدا می شوند محاسبه شود. چه مدت قبل این انشعاب ها جدا شده اند؟ برای پاسخ به این سؤال، باید توالی ها را مقایسه کنید و فاصله ها را اندازه گیری کنید یا به یک سناریوی تکاملی برسید. اما سؤالی که اینجا مطرح می شود اینست که این کار باید روی توالی DNA انجام شود یا پروتئین؟ این سؤال به ظاهر ساده جواب پیچیده ای دارد:

- ✓ **اگر توالی های DNA شما بیش از ۷۰ درصد مشابه اند:**

می توانید یک تطابق چند توالی انجام دهید. اما اگر توالی های شما کد کننده پروتئین هستند این کار پیشنهاد نمی شود.

مشکل تطبیق توالی های DNA اینست که ماتریکس های جایگزینی زیاد با کیفیت نیستند و انجام تطابق ها به خاطر طول توالی ها زمان بیشتری لازم دارد.

- ✓ **اگر توالی های پروتئین شما کمتر از ۷۰ درصد مشابه اند:**

اگر توالی های شما کد کننده پروتئین هستند، بهتر است تطابق را روی پروتئین انجام دهید.

اگر توالی های DNA شما در سطح پروتئین بسیار مشابه باشند، می توانید تطابق را روی پروتئین ها انجام دهید، برای این کار از برنامه pal2nal استفاده کنید:

coot.embl.de/pal2nal

اگر توالی DNA مربوط به پروتئین های خود را ندارید، می توانید با استفاده از برنامه Protogene توالی DNA مورد نظر را به دست آورید. Protogene یکی از ابزارهای Tcoffee است که در آدرس اینترنتی زیر در دسترس می باشد:

www.tcoffee.org

وقتی پروتئین مورد نظر را پیدا کردید، دو احتمال وجود خواهد داشت:

- **اگر اکثر محل های متناظر [1] جهش متفاوت بودند:** این مورد از نمونه های بارز اشباع شدن است. در این موارد، اندازه گیری فاصله ها در درخت در سطح پروتئین، صحیح تر خواهد بود.
- **اگر محل های متناظر اشباع نبودند:** از نظر تئوری اندازه گیری فاصله ها در سطح DNA صحیح تر خواهد بود.

در عمل، کار کردن در سطح پروتئین راحت تر است، مگر اینکه توالی های شما تقریبا مشابه باشند. صحت کار کردن در سطح پروتئین ممکن است به اندازه کار کردن در سطح DNA نباشد، اما در اغلب موارد می توان نتایج قابل قبولی بدست آورد.

پیراهن سفید، اسپاگتی، تکامل و فیلوژنی!

وقتی به تکامل فکر می کنید، ناخودآگاه ذهن شما به مسائلی مثل انتخاب طبیعی، بقای اصلح، تنازع برای بقا و سایر رفتارهای اینگونه معطوف می شود. این نظریه "داروینیسم جدید [2]" نامیده می شود. این نظریه قدیمی اما پر طرفدار بیان می کند که تنها دلیل حفظ شدن یک موتاسیون در یک جمعیت اینست که این موتاسیون یک صفت مطلوب برای ادامه حیات را به موجود زنده بدهد. اما این نظریه توجیهی برای جهش های متناظر یا دیگر جهش های مضر ندارد.

[1] synonymous
[2] Neo-Darwinism

در دهه ۱۹۶۰، افزایش امکان دسترسی به اطلاعات مولکولی، سبب ایجاد نظریه نوترالیسم [1] شد. نوترالیسم که تئوری نوترال هم نامیده می شود، با بیان این مطلب که اکثر جهش ها کشنده یا خنثی بوده اند، نظریه داروینیسم را زیر سؤال برد. کاملا واضح است که جهش های کشنده فورا از بین می روند، و از دیدگاه نظریه نوترالیسم، احتمال ایجاد یک جهش غیر کشنده در یک جمعیت و گسترش یا عدم گسترش آن اغلب بستگی به شانس دارد. حتی اگر این جهش ها مفید باشند، شانس عامل اصلی در حفظ یا بقای آنهاست. توجه کنید که جهش های خنثی، الزاما از نوع جهش های خاموش [2] نیستند، بلکه بعضی از این جهش ها باعث تغییرات جزئی در پروتئین ها می شوند، بدون اینکه در عملکرد پروتئین ها تغییری ایجاد نمایند. البته بعد از اینکه این جهش های غیر مضر و غیر مفید تثبیت شدند می توانند جایگاه ویژه ای پیدا کنند، و فوایدی به فرد حامل خود برسانند. نوترالیسم یک تئوری بسیار قوی است، چون بیان می کند که چگونه گونه های جانداران پتانسیل تطابق را قبل از اینکه به آن نیاز داشته باشند، بدست می آورند.

در استفاده از فیلوژنی، این جهش های خنثی بسیار جالب توجه هستند. اما الان هدف ما بررسی تطابق نیست، بلکه ما می خواهیم فاصله های تکاملی را اندازه گیری کنیم و بهترین راه این اندازه گیری اینست که تعداد حوادث اتفاقی و شبه خنثی [3] که برای توالی های شما از زمان جدا شدن از جد مشترکشان رخ داده است را محاسبه کنید. این جهش ها را می توان شبیه هر تیک ساعت در نظر گرفت که مدت زیادی کار می کرده است (یک جهش در هر زمان، برای یک میلیون سال).

برای درک بهتر مطلب فرض کنید امروز صبح، مثل روزهای دیگر، دو فرزند شما دو پیراهن سفید پوشیده اند و بیرون رفته اند. در طول روز بچه های شما اسپاگتی چرب و پر سسی را خورده اند. اکنون فرض کنید هر نقطه رنگی روی پیراهن بچه ها یک جهش است و تعداد نقطه ها نشاندهنده این است که چه مقدار از روز گذشته است. حال اگر می خواهید بدانید برای چه مدتی پیراهن ها تن بچه ها بوده، باید تعداد نقطه ها را بشمارید. این یک مثال از نوترالیسم است و این شرایط، شرایط ایده آل برای مقایسه توالی هاست.

حال فرض کنید کمی انتخاب اتفاق می افتد. آن دو از اینکه شما ممکن است چه واکنشی بروز دهید، بسیار مضطرب هستند بنابراین پیراهنهای خود را تا جایی که می توانند تمیز می کنند. این موضوع شبیه جهش های انتخابی است و شما نمی توانید بفهمید آنها تا چه مدت پیراهنها را پوشیده اند. جهشهای انتخاب نشده (خنثی)، برای ارزیابی فاصله زمانی بسیار مناسب تر هستند، زیرا به تدریج تجمع می یابند. در DNA کد کننده، جهش های انتخاب نشده بیشتر متداول هستند. به همین خاطر است که DNA برای اندازه گیری فاصله زمانی مناسب تر از توالی های پروتئینی است که تکامل آنها محدود تر می باشد.

حالا تصور کنید که بچه ها برای دومین بار و سومین بار هم اسپاگتی پر سس خورده اند، لکه های جدید روی لکه های قدیمی را پوشانده است و پیراهن آنها از لکه های اسپاگتی اشباع شده است. بدین ترتیب حدس زدن تعداد دفعاتی که آنها اسپاگتی خورده اند، بسیار مشکل است. به همین ترتیب هنگامیکه یک یا دو جهش روی هر مکان خنثای توالی DNA ما اتفاق می افتد، تعیین اینکه تا چه مدتی این اینکار اتفاق می افتاده است مشکل می گردد. هرچه این اشباع بیشتر باشد، امکان این سنجش نامحتمل تر می گردد.

حال فرض کنید یکی از این کودک با لباس کاملا تمیز و دیگری با لباس گلی به خانه وارد شوند. این موضوع بیش از آنکه نشان دهنده آن باشد که هرکدام چه مدتی پیراهن هایشان را برتن داشته اند، می تواند بیانگر آن باشد که آن دو در دو محل متفاوت بوده اند یا دو گونه رفتار مختلف داشته اند. چنین تفاوتهایی در مورد ژنها نیز اتفاق می افتد و گرچه ما مایل هستیم که فکر کنیم یک ساعت مولکولی جهانی وجود دارد، اما می دانیم که این ساعت برای هر ژن با سرعت متفاوتی تیک تاک می کند.

انتخاب توالی ها برای ساخت درخت ژنی یا درخت گونه ها

ژنهای همولوگ [4] ژنهایی هستند که از یک جد مشترک منشا می گیرند. ارتباط این ژن ها می تواند به سه صورت باشد:

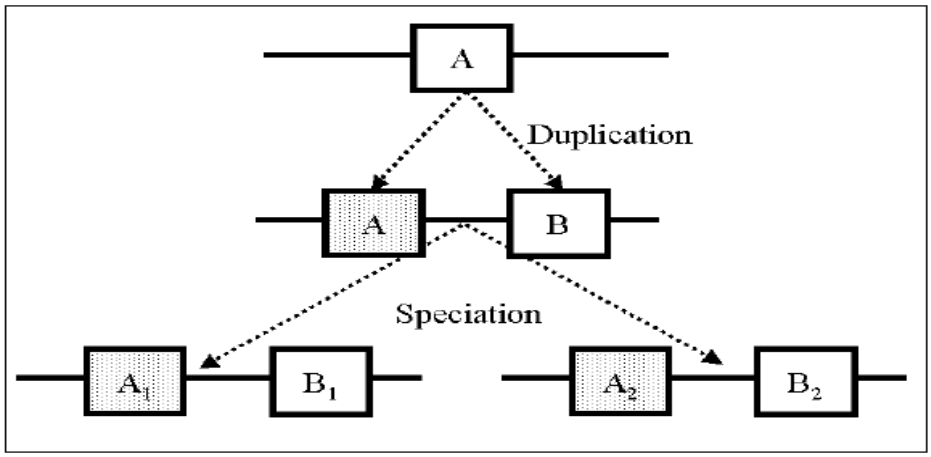
- ✓ **اورتولوگ:** این ژنها طی پدیده جدایی گونه ها از هم جدا شده اند. جدایی گونه ها پدیده ای است که طی آن یک گونه به دو زیر گروه تقسیم می شود که در طول زمان ترکیب ژنتیکی آن ها آنقدر تغییر می کند که به دو گونه مجزا تبدیل می

[1] Neutralism
[2] Silent mutation
[3] Quasi-neutral
[4] homologous

شوند. با فرض اینکه در ژنوم دو گونه حاصل نوآرایی رخ ندهد، دو ژن که از یک ژن خاص در ژنوم جد مشترک منشا گرفته باشند، اورتولوگ نامیده می شوند. زیست شناسان معمولا انتظار دارند دو ژن اورتولوگ عملکرد و ساختار یکسان داشته باشند. در شکل ۱-۱۳، A1 و A2 با هم و B1 و B2 با هم اورتولوگ هستند.

- ✓ **پارالوگ:** ژنهای همولوگی هستند که طی پدیده دو برابر شدن از هم جدا شده اند. در پدیده دو برابر شدن در یک ژنوم، یک ژن دو برابر می شود. یکی از ژن های حاصل ممکن است عملکرد اولیه خود را حفظ کند و دیگری ممکن است عملکرد جدیدی بدست آورده باشد. بنابراین می توان انتظار داشت که ژنهای پارالوگ عملکردهای متفاوت اما مرتبط داشته باشند. به عنوان مثال، در شکل ۱-۱۳، A1 و B1 پارالوگ هستند.
- ✓ **زنولوگ** [1] : Xeno یک واژه یونانی به معنی "خارجی" است. ژن های زنولوگ حاصل یک انتقال عرضی بین دو موجود زنده یا به عبارتی انتقال مستقیم DNA بین دو گونه هستند. در واقع یک ژن وارد ژنوم دیگری می شود که تاریخچه متفاوتی از آن دارد. یک نمونه بارز از انتقال عرضی، بدست آوردن ژن ایزولوسیل-tRNA سنتاز از موجود میزبان، توسط تعدادی از باکتری هاست. ایزولوسیل-tRNA سنتاز پروتئینی است که در سنتز دیگر پروتئین ها نقش دارد، و به نظر می رسد که بدست آوردن آن توسط باکتری ها، به مقاوم شدن آن ها در برابر آنتی بیوتیک ها کمک می کند. هنگامی که این اتفاق می افتد، ایزولوسیل-tRNA سنتاز جدید، زنولوگ دیگر tRNA سنتازهای موجود در باکتری خواهد بود.

زمانی که برای ساخت یک درخت فیلوژنتیک یک گروه از ژن های همولوگ را انتخاب می کنید، به اصطلاح زیست شناسان در واقع یک درخت ژنی درست کرده اید. این درخت تاریخ تکاملی ژنهایی که در آن وجود دارند را نشان می دهد.



شکل ۱-۱۳: ارتباط اورتولوگ و پارالوگ

زمانیکه تمام اعضای پارالوگ یک خانواده بزرگ ژنی انسانی را انتخاب می کنید، درخت ژنی حاصل فقط تاریخ تکاملی این ژن ها را نشان می دهد. همچنین شما از این درخت فقط می توانید برای بازسازی زنجیر دوبرابر شدن هایی که روی یک ژن منفرد انجام شده تا به شرایط کنونی منجر شده استفاده کنید.

زمانیکه یک گروه از ژن های اورتولوگ از گونه های مختلف را انتخاب می کنید، درخت حاصل بسیار شبیه درخت گونه ها [2] می شود، و می توانید با استفاده از آن جدا شدن گونه ها را بازسازی کنید. بهترین مثال از این نوع درخت ژنی، درخت

[1] Xenologous
[2] Species tree

فیلوژنتیک RNA ریبوزومی است که زیست شناسان با استفاده از آن درخت بزرگ زندگی را بازسازی می کنند. ژن های RNA ریبوزومی در همه گونه ها وجود دارند و در گونه های مختلف نسبت به هم اورتولوگ هستند.

البته شما همچنین می توانید بررسی خود را برعکس انجام دهید، یعنی از روی درخت تشخیص دهید که آیا توالی های مورد نظرتان اورتولوگ هستند یا پارالوگ. این روش زمانی معتبر است که شما "درخت گونه ها" را به خوبی بشناسید.

دو ژن همولوگ از دو گونه متفاوت لزوما اورتولوگ نیستند. این ژنها ممکن است پارالوگ باشند (مثل A1 و B2 در شکل ۱-۱۳). اشتباه کردن ژنهای پارالوگ و اورتولوگ در درخت های فیلوژنتیک مهمترین مشکل در کار کردن با این درخت هاست و متأسفانه اجتناب از آن هم مشکل است.

هیچ راه حل تئوری برای تایید اورتولوگ بودن دو ژن وجود ندارد و اثبات آن هم عملا غیر ممکن است. اما در عمل استفاده از معیار های تجربی نتایج خوبی داشته است. در این قسمت استفاده از BLAST را در این زمینه تمرین می کنیم:

۱. از ژنوم A یک توالی A انتخاب کنید.
۲. در BLAST یک جستجو انجام دهید و توالی A را با تمام توالی های موجود در ژنوم B مقایسه کنید (چگونگی جستجو در BLAST را در فصل ۷ می توانید ببینید.).

اولین نتیجه ای که برای این جستجو بدست می آید، توالی B است.

۳. در BLAST یک جستجو انجام دهید و توالی B را با تمام توالی های موجود در ژنوم A مقایسه کنید.

اولین نتیجه این جستجو توالی A خواهد بود، به این ترتیب شما با توالی A و توالی B مثل دو توالی اورتولوگ از دو ژنوم A و B رفتار کنید.

این نتایج نشاندهنده این نیست که توالی A و B واقعا اورتولوگ هستند، اما می تواند بیانگر این باشد که اگر آنها را اورتولوگ بدانید، زیاد هم اشتباه نکرده اید.

این روش تنها زمانی درست خواهد بود که دو ژنوم مورد نظر شما *کامل* باشند و اندازه پروتئوم [1] حاصل از آنها هم تقریبا برابر باشد.

Eugene Koonin یک مجموعه بسیار با کیفیت از ژن های اورتولوگ را در NCBI جمع آوری کرده است، و برای تشخیص اورتولوگ بودن آن ها از همین معیار، اما به شکلی کمی پیچیده تر، استفاده کرده است. این مجموعه تحت عنوان COG [2] شناخته می شود و شما می توانید آنرا در آدرس اینترنتی www.Ncbi.nlm.nih.gov/COG ببینید. مجموعه های دیگری از ژنهای همولوگ HOGENOM (که قبلا با نام HOBACGEN شناخته می شد) و HOVERGEN هستند که توسط Pole Bioinformatique Lyonnais ایجاد شده اند. (http://pbil.univ-lyonl.fr/)

ایجاد یک مجموعه کامل

لغت "کامل" فقط یک شوخی بود، اما در جدول ۱-۱۳ چند نکته که برای جمع آوری یک مجموعه از ژنهای اورتولوگ لازم است مد نظر قرار دهید، آورده شده است.

[1] Proteome
[2] Clusters of Orthologous Groups

Table 13-1 Preparing a Set of Sequences for Making a Phylogenetic Tree

Problem	*Reason and Solution*
Avoid sequence fragments	Incomplete sequences make multiple sequence alignments and tree reconstruction methods very sick! You want to avoid these at all costs — or you at least want to use the same fragment for all the sequences.
Avoid Xenologs	Unless your purpose is to study them, avoid including genes that result from lateral transfer.
Avoid recombinant sequences	Some proteins result from the combination of several proteins. This is especially common in viruses. In terms of phylogeny, such proteins have two ancestors rather than one — and standard tree methods are not equipped to represent this kind of relationship.
Avoid large complex families	Very large families that contain various domains and repeats can be very tricky to analyze. The ABC transporter is a good example of this kind of troubled family. If you can, stay away from these protein gangs — or try to work on smaller, more uniform subsets.
Try to make a small set	Phylogenetic reconstruction is one of the only cases where you're allowed to align very similar sequences — as many as you want. Unfortunately, it isn't because you need it that you can do it. If you've been through Chapter 9, you know this is not a good thing because most methods find it hard to deal with large datasets. Analyzing large sets means that you may not be able to use the most powerful methods, or that you will need to install these programs on your computer.
Add an outgroup to your dataset	An *outgroup* is a sequence that you know has diverged long ago from the rest of the set. For instance, horse hemoglobin would be an outgroup if you're analyzing primate hemoglobin. This outgroup helps you by adding a root to your tree that symbolizes the first common ancestor of all your sequences. The choice of an outgroup relies mostly on biological criteria, yet you must also make sure that this sequence is similar enough to the rest of the set so your multiple-sequence-alignment method can align it (see Chapter 9).

فسیل های زنده وجود ندارند! بجز در مواردی که از توالی های DNA فسیل استفاده می کنید، تمام توالی هایی که شما مقایسه می کنید سن یکسانی دارند. حتی اگر یکی از گونه ها تشابه خود با جد مشترک را بیشتر از بقیه گونه ها حفظ کرده باشد، این گونه هم به اندازه پیچیده ترین اعضای این خانواده، متحمل جهش های خنثی شده است.

آماده شدن برای انجام تطابق چند توالی

بسیاری از روش های بازسازی درخت های فیلوژنتیک نیاز به یک جدول بخصوص به نام *ماتریکس فاصله*[1] دارند. این جدول حاوی فاصله ها یا تعداد وقایع تکاملی اتفاق افتاده بین هر جفت توالی در مجموعه شما است.

از منظر تئوری، شما می توانید با مقایسه دو به دو[2]ی تمام توالی ها این ماتریکس را بسازید. اما این روش در عمل کارساز نیست. مقایسه دو به دو به تنهایی کافی نیست. (برای اطلاعات بیشتر در مورد مقایسه دو به دو به فصل ۸ مراجعه کنید) تنها راه اندازه گیری صحیح فاصله های بین توالی ها یک فرایند دو مرحله ای است:

۱. انجام یک "تطابق چند توالی" با استفاده از دو توالی.
۲. اندازه گیری فاصله ها با استفاده از "*تجسم دو به دوی توالی ها*"[3]

تجسم دو به دو، در واقع یک تطابق دو به دو است که در خلال تطابق چند توالی انجام می شود. این تطابق با بهره گیری از اطلاعات موجود در **همه** توالی های دیگر انجام می گیرد.

کیفیت تطابق چند توالی عامل محدود کننده واقعی در ساخت یک درخت است. با یک تطابق کم کیفیت به هیچ وجه نمی توان یک درخت خوب ساخت. (اگر با تطابق چند توالی آشنا نیستی، به فصل ۹ مراجعه کنید.)

انجام تطابق چند توالی

برای این کار می توانید از **هر کدام** از روش های معرفی شده در فصل ۹ استفاده کنید. این روش ها عبارتند از:

- ✓ **ClustalW :** www.ebi.ac.uk/clustalw
- ✓ **MUSCLE:** phylogenomics.berkeley.edu/cgibin/muscle/input_muscle.py
- ✓ **Tcoffee :** www.tcoffee.org

اطمینان از انتخاب تطابق چند توالی درست

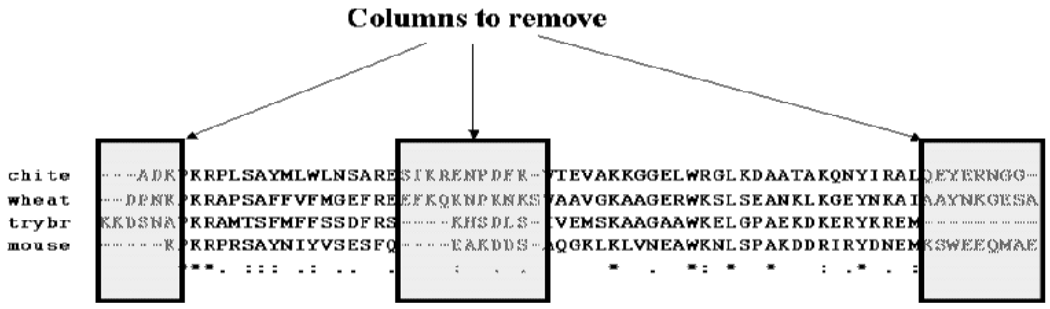
قبل از استفاده از تطابق چند توالی برای ساخت یک درخت، باید مطمئن شوید که این تطابق به اندازه کافی صحیح است. در این بخش چند معیار و لیستی از کارهایی که باید انجام بدهید تا کیفیت کار خود را بالا ببرید، به شما معرفی می کنیم. (شکل ۲-۱۳ انواع نواحی که در مراحل ذکر شده در ادامه، مورد استفاده قرار می گیرند را نشان داده است.)

برای اطمینان از اینکه تطابق چند توالی شما هم صحیح و هم مناسب است، مراحل زیر را دنبال کنید:

۱. اطمینان حاصل کنید که تا حد ممکن ستون های بدون شکاف[4] انتخاب کرده اید:

شکاف ها در اغلب روش های بازسازی درخت های فیلوژنتیکی ایجاد اختلال می کنند. بعضی از این روش ها مثل ClustalW از *تکنیک های حذف کامل*[5] استفاده می کند و هر ستونی را که دارای شکاف باشد، نادیده می گیرد.

[1] Distance matrix
[2] Pairwise comparison
[3] Sequences' pairwise projection
[4] Gap-free columns
[5] Complete-deletion techniques



شکل ۲-۱۳: آماده کردن یک تطابق چند توالی

۲. انتها ها را از برنامه تطابق خارج کنید.

انتهاهای آمینی و انتهاهای کربوکسیلی بخوبی حفاظت شده نیستند و بنابراین بخوبی تطابق داده نمی شوند. شما می توانید آنها را حذف کنید (شکل ۲-۱۳).

۳. مناطق غنی از شکاف را از برنامه خارج کنید.

نواحی داخلی غنی از شکاف در برنامه تطابق توالی معمولا نواحی هستند که تشکیل حلقه می دهند. در صورت وجود این نواحی، حتی اگر برنامه شما تطابق را انجام دهد، نمی توان با اطمینان گفت که تطابق درست انجام شده است.

۴. مطمئن شوید که قطعه های حاوی بیشترین اطلاعات را حفظ کرده اید.

تطابق چند توالی ایده آل برای ساخت یک درخت، تطابق با کیفیت توالی های با تشابه کم است، به طوریکه هر جایگاه حاوی یک اثر از تاریخچه خانوادگی باشد.

چند نکته در این مورد:

- در فصل ۹، خصوصیات یک قطعه خوب در تطابق چند توالی را ذکر کردیم؛ چنین قطعه ای معمولا به طول ۳۰-۲۰ اسید آمینه است، و تعدادی جایگاه های محافظت شده دارد.
- در صورت تمایل، می توانید با استفاده از Tcoffee تطابق چند توالی را ارزیابی کنید و ستون هایی را که احتمال دارد به درستی آنالیز نشده اند را حذف کنید. (چگونگی انجام این کار را می توانید در فصل ۹ ببینید.)
- بهترین راه برای ویرایش تطابق چند توالی انجام شده، استفاده از Jalview است. این برنامه در فصل ۱۰ معرفی خواهد شد.

به عنوان یک راه حل ساده و آسان اما نه خیلی مطمئن، می توانید از برنامه Microsoft Word برای حذف قطعات مورد نظر استفاده کنید:

۱. دکمه Alt را نگه دارید و با موس ستون های مورد نظر را انتخاب کنید.

۲. دکمه Delete را فشار دهید.

هرگز سعی نکنید با استفاده از برنامه Microsoft Word تغییرات دیگری غیر از آنچه در بالا ذکر شده، روی توالی ها و تطابق ها انجام دهید.اگر توالی های شما نیاز به ویرایش دارند، از یک ویرایش گر تطابق ها مثل Jalview استفاده کنید.

ساخت درخت مورد نیاز شما

روش های ساخت یک درخت به سه دسته عمده تقسیم می شوند: روش هایی که اساس آن ها فاصله است (روش های فاصله[1])، روش هایی که اساس آنها صرفه جویی است (روش های صرفه جویی[2])، و روش هایی که بر اساس احتمالات کار می کنند (روش های احتمالات[3]). هیچ کدام از این روش ها کامل نیستند، و هر کدام نقاط ضعف و قوت خاص خود را دارند.

در این فصل نحوه ساخت درخت را، با استفاده از روش های "فاصله" و "بیشترین احتمال"، یاد می گیرید. درخت های ساخته شده با روش فاصله لزوما صحیح ترین درخت ها نیستند، اما بطور کلی این درخت ها کاربردی هستند و کار با آنها آسان است. شما از این روش در اکثر مواقع می توانید استفاده کنید، و غالبا با این روش می توان به جواب های بسیار منطقی دست یافت. روش های حداکثر صرفه جویی و حداکثر احتمال از نظر تئوری صحیح ترین روش ها هستند، اما برای انجام آن ها زمان بیشتری هم لازم است. به همین خاطر متخصصین روش بیشترین احتمال را با احتیاط به کار می گیرند و در مورد بهترین روش هنوز بحث وجود دارد.

زمانی که تصمیم گرفتید از روش فاصله استفاده کنید، به ۴ جزء لازم برای این کار را در نظر بگیرید:

- ✓ **تطابق چند توالی:** از چه روشی؟ چه توالی هایی؟
- ✓ **اندازه گیری فاصله ها:** چه نوع ماتریکس جایگزینی؟
- ✓ **الگوریتم بازسازی درخت:** UPGMA، NJ، Fitch یا Kitsch
- ✓ **نوع درختی که می خواهید نمایش بدهید؟** یک درخت بدون نشان دادن فاصله ها (کلادوگرام[4])، یا درختی که فاصله ها را نیز نشان می دهد (فنوگرام[5]) ؟ یک درخت بدون ریشه یا با ریشه؟

در این بخش یاد می گیرید که چگونه با استفاده از اینترنت یک درخت طراحی کنید و آنرا به هر فرمتی که می خواهید نمایش دهید.

طراحی درخت مورد نظر شما

در ساخت درخت فیلوژنتیک مانند آشپزی باید این توانایی را داشته باشید که اجزاء متشکله را به درستی با هم مخلوط نمایید. اما در ساخت درخت فیلوژنتیک، بر خلاف آشپزی، یک دستورالعمل ساخت همیشه کارساز نیست؛ در حقیقت کیفیت درخت، بیشتر به سلیقه و روحیه شما بستگی دارد.

مانند بعضی غذا ها، برای ساخت درخت فیلوژنتیکی نیز چندین راه وجود دارد. اولین راه استفاده از ClustalW است؛ این روش سریع، بدون پیچیدگی، و تا حدودی شبیه درست کردن غذاهای آماده است. در ادامه به شما یاد می دهیم که چگونه به راحتی با cut و paste کردن توالی های خود، بدون برخورد با هیچگونه مشکلی، یک درخت بسازید.

پس از این قسمت، دستورالعمل دیگری به شما می دهیم که بیشتر مناسب افرادی است که دوست دارند از بازار سبزی تازه بخرند، و خودشان سالاد درست کنند. در این روش شما می توانید همه اجزاء به کار رفته در ساخت درخت را تحت کنترل خود داشته باشید. افراد خیلی دقیق و حساس بهتر است از این روش استفاده کنند. این دستورالعمل Phylip نامیده می

[1] distance methods
[2] parsimony methods
[3] likelihood methods
[4] a cladogram
[5] a phenogram

شود. این روش بسیار مهم است، و متخصصین فیلوژنی معروف جهان از آن استفاده می کنند. اما نگران نباشید، روش Phylip بسیار ساده است.

ساختن سریع یک درخت فیلوژنتیک با استفاده از ClustalW

اگر به یک درخت نیاز دارید ولی به هیچ وجه از باغبانی خوشتان نمی آید، بهترین راه حل برای شما استفاده از ClustalW است. کار اصلی ClustalW تطابق چند توالی است، اما بعضی سرور های امکان ساخت درخت فیلوژنتک را برای شما فراهم می کنند.

سرور EBI از مجموعه ClustalW (و نیز در اغلب سرور های دیگر ClustalW)، اگر از توالی های تطابق نیافته استفاده کنید، این برنامه به شما یک درخت تحویل خواهد داد. اما این درخت، یک درخت فیلوژنتیک نخواهد بود؛ در واقع درخت حاصل یک درخت راهنما است که ClustalW از آن برای تطابق چند توالی استفاده می کند. بطور کلی، هر درختی با پسوند .dnd که در ClustalW به دست می آید، یک درخت فیلوژنتیک نیست.

شما نمی توانید از این گونه درخت ها به عنوان یک درخت فیلوژنتیک استفاده کنید. در واقع این کار یک اشتباه خواهد بود و نتایج غیر صحیحی از آن به دست خواهد آمد.

در این بخش یاد می گیرید که چگونه از سرور EBI برای ساخت درخت فیلوژنتیک استفاده کنید. فرض می کنیم شما قبلا تطابق چند توالی را انجام داده اید، اما اگر این کار را نکرده اید می توانید یک تطابق چند توالی آماده از آدرس اینترنتی زیر دانلود کنید:

www.tcoffee.org/dummy_aln2.html

به عنوان مثال تطابق چند توالی آماده ما حاوی قطعات همولوگ از گاما فیبرینوژن است. این توالی ها از گونه های مختلف، برای بررسی های مقایسه ای، ردیف شده اند. بجای توالی کامل می توانید از قطعاتی از آن استفاده کنید، چرا که این قطعات در تمام طولشان همولوگ هستند. ما به هر کدام از این قطعات نام یک گونه را می دهیم، چون درخت هایی که با نام گونه ها طراحی شده باشند، راحت تر خوانده می شوند(مگر اینکه شما یک حافظه بسیار قوی داشته باشید).

O12957 (Sheep), O02 72 (Moose), O02 3 (Giraffe)

O02 90 (Chevrotain), O02 1 (Beluga) O02 7 (Sperm_whale)

O02 73 (Rorqual), O02 (Pig), O12959 (Peccary)

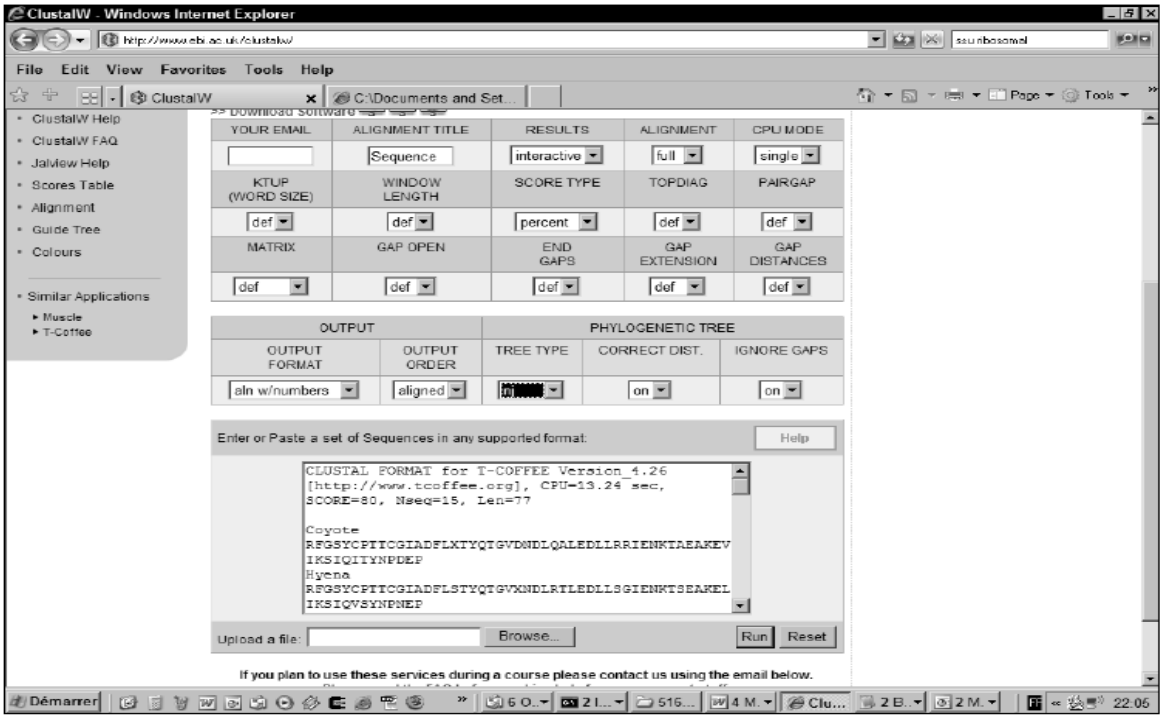
O02 77 (Dromedary), O02 9 (Tapir), O02 2 (Horse)

O02 7 (Hyena), O02 0 (Coyote), O12954 (Hippopotamus)

به هر حال شما می توانید از هر مجموعه ای از توالی های تطابق یافته با یکی از سرورهایی که در فصل ۹ معرفی شدند، استفاده کنید.

۱. به آدرس اینترنتی www.ebi.ac.uk/clustalw/ وارد شوید.
۲. چند توالی تطابق یافته را به پنجره Sequence وارد کنید (شکل ۳-۱۳).

در این قسمت توالی ها را به هر فرمتی که بخواهید می توانید وارد کنید، مثلا فرمت ALN (فرمت ClustalW)، MSF، یا FASTA. در صورتی که از ALN استفاده می کنید، باید توجه کنید که حتما خط اول را که با CLASTAL W ... آغاز می شود را نیز وارد کنید.



شکل ۳-۱۳: طراحی یک درخت فیلوژنتیک با استفاده از ClustalW

۳. در قسمت Phylogenetic Tree، از منوی Tree Type، گزینه NJ را انتخاب نمایید(شکل ۴-۱۳).

در این قسمت شما از بین چندین روش طراحی درخت، یکی را انتخاب می کنید. به طور کلی متخصصین معتقدند که روش NJ[1]، برای کاری که ما می خواهیم انجام بدهیم، بهترین روش است.

زمانیکه گزینه های مربوط به طراحی درخت را تنظیم نمودید (در واقع زمانی که در تمام منوها گزینه ای غیر از None را انتخاب نمودید)، باید توالی های تطابق یافته را در اختیار ClustalW قرار دهید. شاید بپرسید چرا؟ چون زمانیکه تمام گزینه ها تنظیم شده باشند، ClustalW آنالیز تطابق ها را متوقف می کند، و فرض می کند که شما به آن توالی های تطابق یافته داده اید. اگر شما مطابق این فرض عمل نکرده باشید، یعنی به سرور توالی های غیر تطابق یافته داده باشید، ClustalW بر اساس آن ها برای شما یک درخت طراحی می کند. واضح است که درخت حاصل اشتباه خواهد بود (مگر زمانیکه توالی های داده شده بدون شکاف، قابل تطابق باشند).

۴. در قسمت Correct Dist گزینه On را انتخاب کنید.

بدین ترتیب شما قادر خواهید بود جایگزینی های چندگانه (Multiple Substitutions) را اصلاح نمایید. این کار شاخه های درخت شما را بلند تر می کند، و بخصوص برای توالی هایی که خویشاوندی دور با هم دارند، مناسب می باشد.

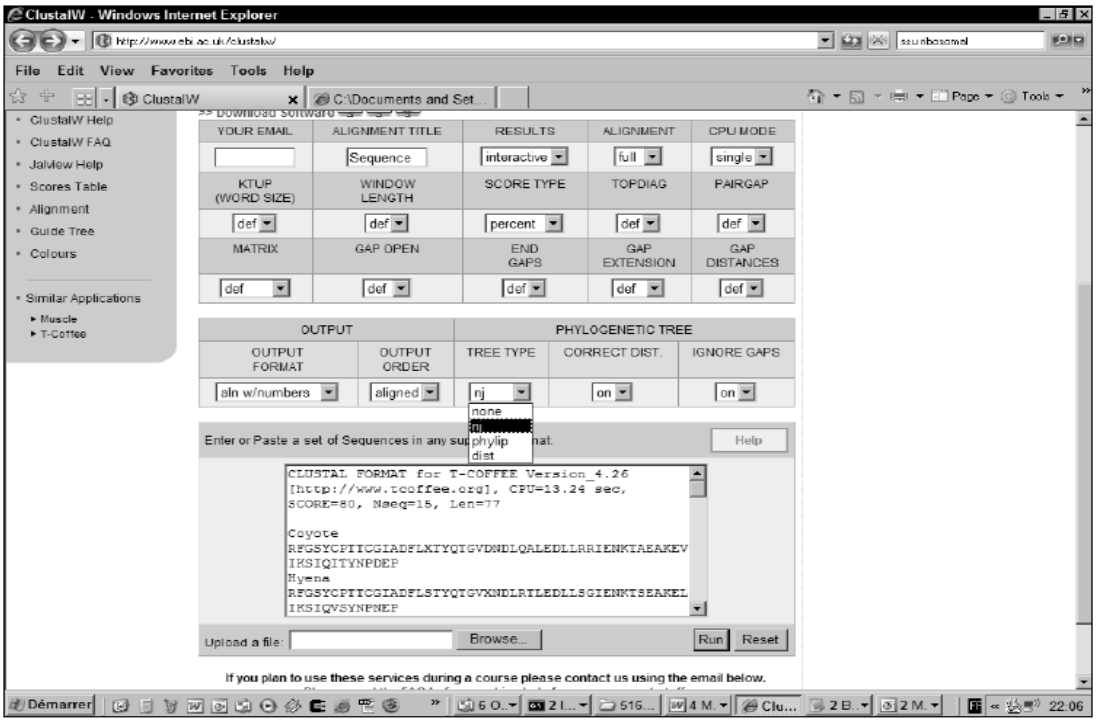
۵. از منوی Ignore Gaps گزینه On را انتخاب کنید.

با این کار ClustalW هر ستونی از توالی های تطابق یافته شما را که دارای شکاف باشد، حذف می نماید، و به این ترتیب می توانید مطمئن باشید که همه توالی ها بر اساس تعداد مشابهی از آمبنواسیدها مقایسه می شوند.

۶. روی دکمه Run کلیک نمایید.

منتظر بمانید تا درخت شما ظاهر شود (۵-۱۳). این درخت بسیار صحیح تر از یک درخت راهنما است، و به وضوح روابط ژنتیکی بین اسب آبی و وال را نشان می دهد.

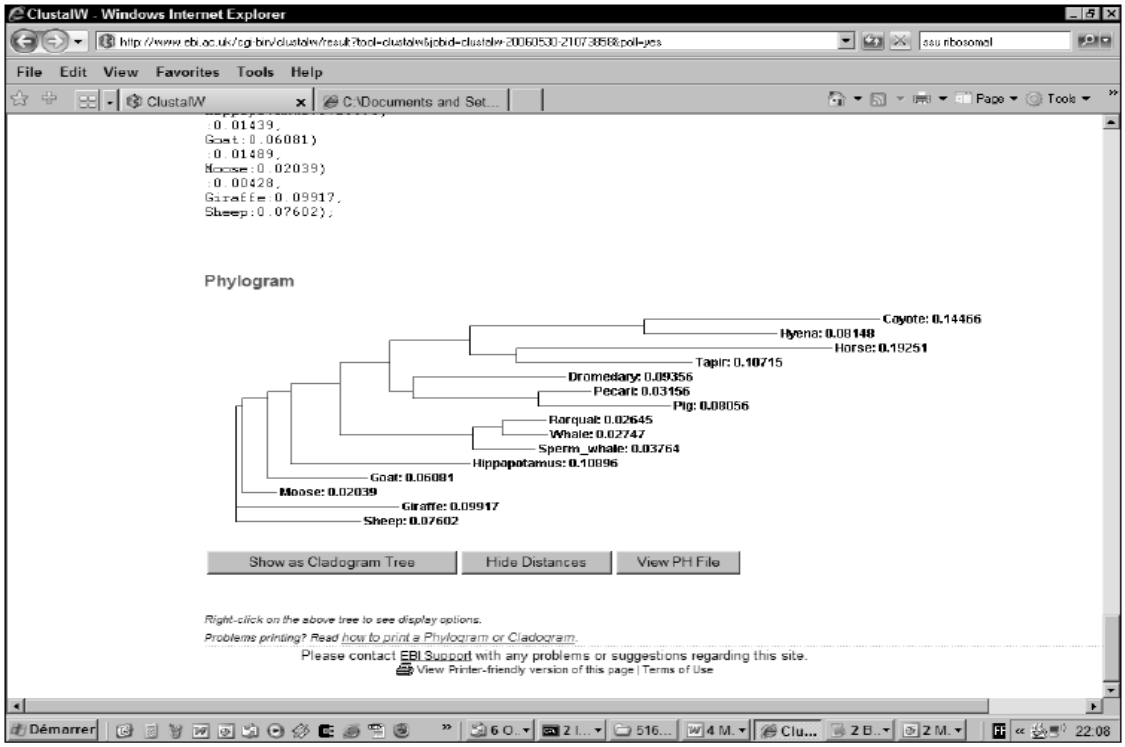
[1] the Neighbor Joining method



شکل ۴-۱۳: درخت فیلوژنتیک: منوی Tree Type

۷. نمایش گرافیکی درخت را ذخیره کنید.

راحت ترین راه برای این کار، گرفتن یک عکس از کل صفحه با استفاده از کلید print-screen (Prt Sc) روی صفحه keyboard است. راه دیگر اینست که تصویر را cut کرده و در محل مورد نظر خود (PowerPoint، Paint و ...) paste نمایید.



شکل ۵-۱۳: یک درخت NJ طراحی شده توسط ClustalW

۸. به پایین صفحه بروید و روی گزینه View PH File کلیک کنید.

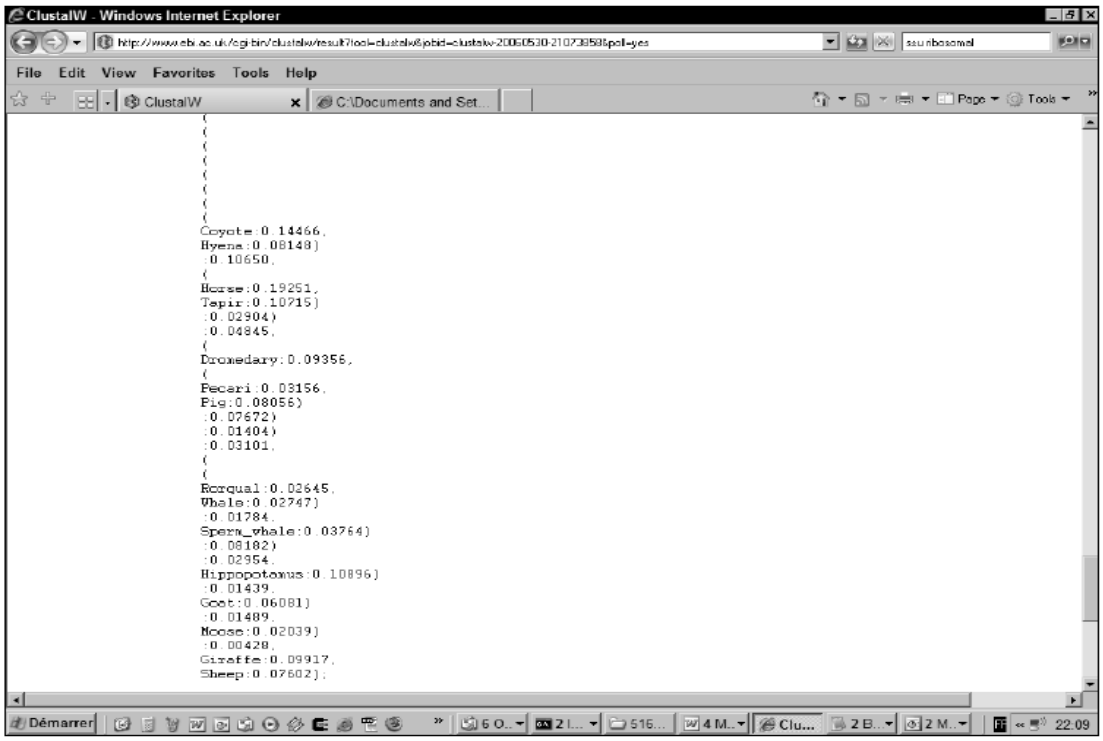
۹. در منوی File با استفاده از گزینه Save As، درخت را با فرمت text ذخیره نمایید.

ClustalW درخت شما را به فرمت Newick text نمایش می دهد (شکل ۶-۱۳). در واقع در این مرحله درخت شما به فرمت Newick text ذخیره خواهد شد.

فرمت Newick text استاندارد ترین روش برای ذخیره یک درخت فیلوژنتیک است، و اغلب برنامه های مرتبط با درخت فیلوژنتیکی از این فرمت استفاده می کنند. اگر می خواهید فقط یک نسخه از درخت خود ذخیره داشته باشید، بهتر است آن نسخه به این فرمت باشد!

ساختن یک درخت با استفاده از Phylip

Phylip یکی از پر استفاده ترین منابع برای طراحی درخت های فیلوژنتیک با صحت بالا است. یکی از خصوصیات خوب Phylip این است که امکان bootstrap کردن درخت را برای شما آسان می نماید (برای توضیح بیشتر به بخش " bootstrapping چیست؟" مراجعه نمایید.). Phylip شما را مطمئن می کند که درخت حاصل به درستی بر اساس تطابق چند توالی است که قبلا انجام داده اید.



شکل ۶-۱۳: فرمت Newick درخت

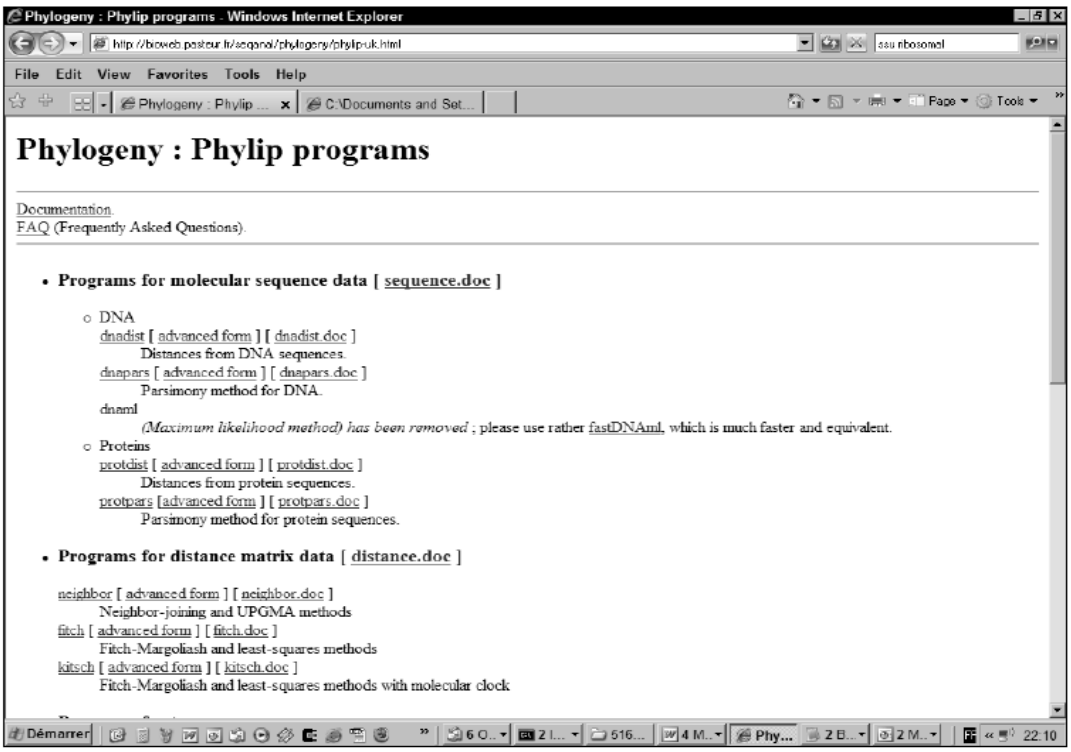
اگر قصد دارید یک متخصص شوید، Phylip تمام امکانات لازم را در اختیارتان قرار می دهد. با استفاده از این برنامه شما می توانید تقریباً هر نوع درختی را با هر روشی که دوست داشته باشید، بسازید. در ادامه مراحل لازم برای ساخت یک درخت فیلوژنتیکی با استفاده از اطلاعات bootstrap ارائه شده است:

قبل از اینکه وارد این سرور شویم، لازم است که یک تطابق چند توالی با کیفیت بصورت آماده داشته باشید. برای این کار یک تطابق چند توالی آماده را می توانید از آدرس اینترنتی www.tcoffee.org/dummy_aln2.html استخراج کنید.

۱. به آدرس اینترنتی bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html وارد شوید.
صفحه Phylogeny: Phylip Programs ظاهر می شود (شکل ۷-۱۳). این صفحه همه برنامه های Phylip که بر روی این سرور نصب شده اند را نشان می دهد. با استفاده از این برنامه ها، یکی پس از دیگری، می توانید یک درخت فیلوژنتیکی درست کنید. تنها نکته ای که باید رعایت کنید اینست که برنامه ها را با ترتیب درست استفاده نمایید.

۲. در زیر سرعنوان Proteins، درست کنار سمت راست لینک protdist، روی لینک advanced form کلیک کنید.
صفحه Phylip: Protdist ظاهر می شود. در این قسمت نکات زیر را باید در نظر بگیرید:

- برای ساخت یک درخت فاصله، لازم است که ابتدا یک ماتریکس فاصله بسازید، که حاوی فاصله توالی ها به صورت دو به دو باشد. این فاصله ها در تطابق چند توالی اندازه گیری می شوند.
- بعضی روش ها این امکان را برای شما فراهم می کند که فاصله های دو به دو را روی DNA یا پروتئین اندازه گیری کنید. برنامه های دیگری نیز وجود دارند که این کار را برای توالی های با خویشاوندی نزدیک یا دور انجام می دهند.
- روش Protdist، روشی است که با استفاده از ماتریکس جایگزینی فاصله بین توالی های تطابق یافته را اندازه گیری می کند.



شکل ۷-۱۳: Phylip به صورت آنلاین در مؤسسه پاستور

- روش Protpars کار مشابهی را انجام می دهد، با این تفاوت که تعداد جهش های در سطح DNA که برای تفاوت های آمینواسیدی مشاهده شده لازم هستند، را نیز مورد ارزیابی قرار می دهد.

۳. در صفحه Phylip: Protdist، در قسمت Your E-Mail، آدرس پست الکترونیک خود را وارد کنید.

باز هم چند نکته در مورد این قسمت:

- اگر سرور گرفتار باشد یا زمانی که یک تطابق بزرگ را می خواهید آنالیز کنید، Phylip به جای اینکه نتایج را به صورت آنلاین به شما نشان دهد، آن ها را از طریق پست الکترونیک برای شما ارسال می کند.
- در این صورت نامه الکترونیکی که شما دریافت می کنید، حاوی یک فایل URL است. زمانیکه این فایل را باز کنید، صفحه ای مشابه شکل ۱۰-۱۳ ظاهر می شود.
- مواظب باشید که آدرس پست الکترونیک قلابی به سرور ندهید! در این صورت هیچ وقت درختی را که ساخته اید نخواهید دید!

۴. توالی های تطابق یافته را در پنجره Sequence وارد کنید (شکل ۸-۱۳).

فراموش نکنیدکه خط سرعنوان را که با CLUSTALW شروع می شود، همراه با توالی کپی کنید.

۵. روی لینک Bootstrap options کلیک کنید.

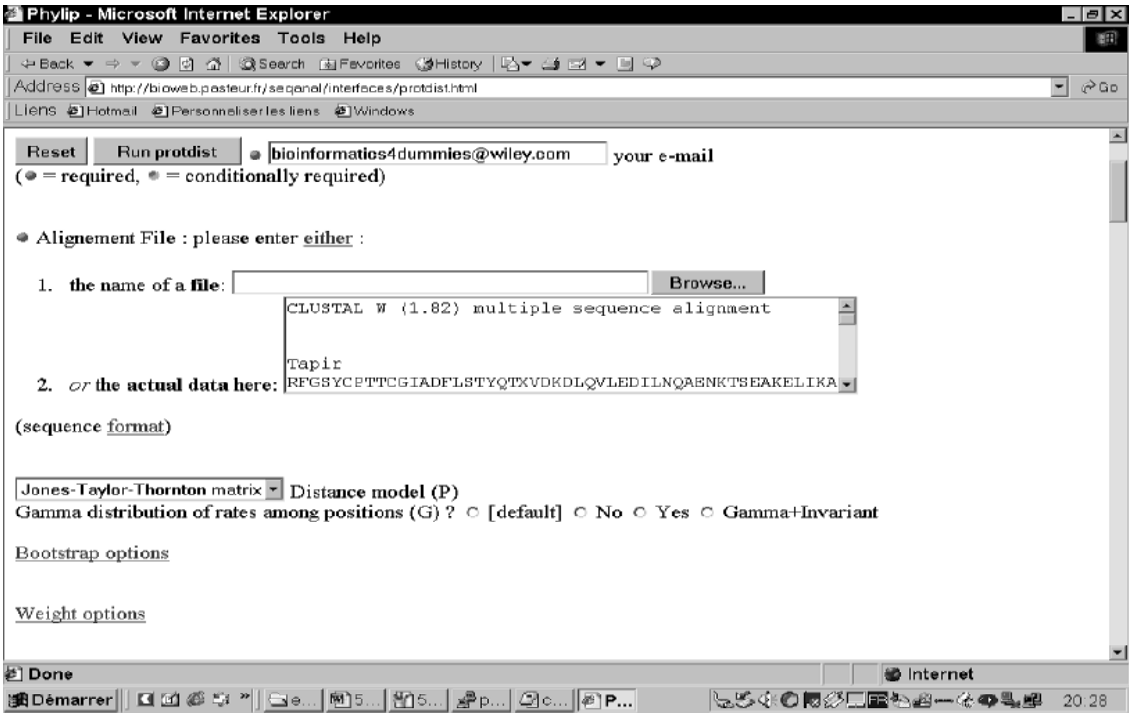
با این کار به قسمت Bootstrap Options وارد شده اید.

۶. مربع کنار گزینه Perform a Bootstrap Before Analysis را تیک بزنید (۹-۱۳).

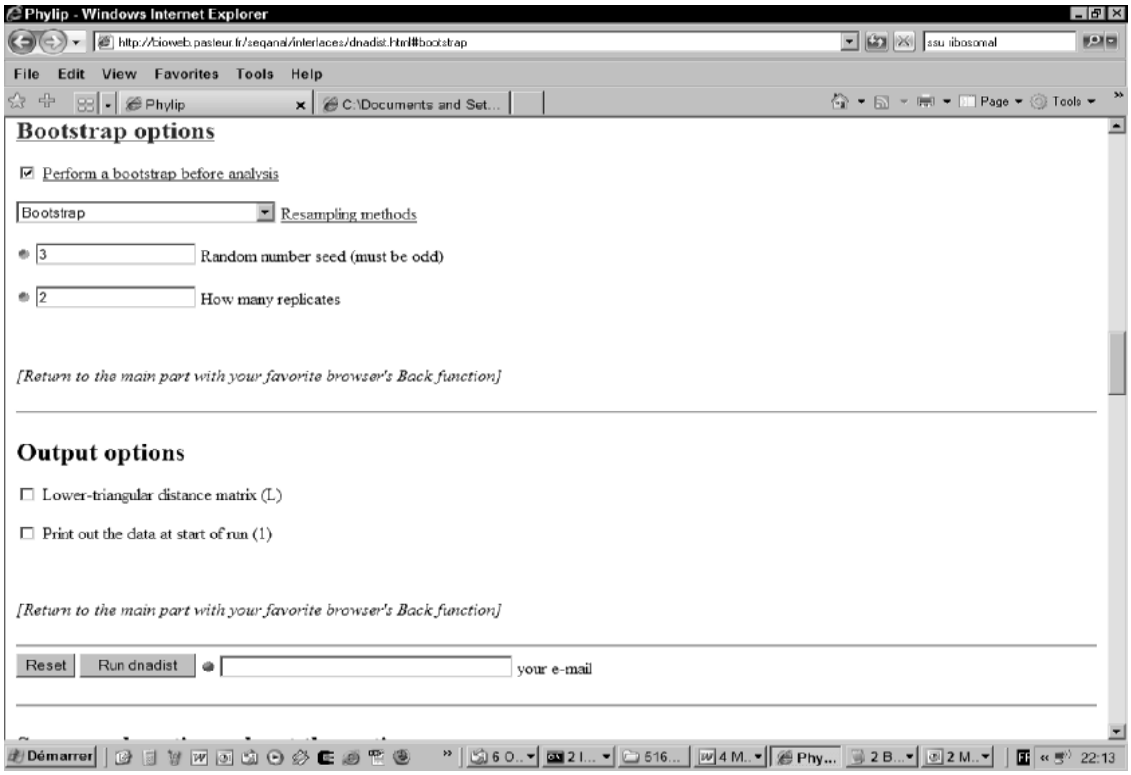
شما می توانید با Bootstrap کردن داده هایتان کیفیت درخت فیلوژنتیکی خود را ارزیابی کنید.

۷. یک عدد فرد، مثلا ۱ یا ۳، در قسمت Random number seed وارد کنید.

بخش seed تولید عددهای فرد را طی عمل Bootstraping کنترل می نماید. عددهای مختلف در این قسمت نتایج متفاوتی ایجاد می کند.



شکل ۸-۱۳: سرور Protdist



شکل ۹-۱۳: گزینه های مربوط به Bootstrap در protdist

۸. تعداد تکرار ها[1] را حداقل ۱۰۰ قرار دهید.

تعداد تکرار ها در واقع تعداد دوره های Bootstrap است که شما انجام می دهید.

شما قسمت های بعدی نیاز دارید این عدد را دوباره وارد کنید، پس آنرا به یاد داشته باشید!

برای یک آنالیز واقعی شما حداقل به ۱۰۰ تکرار نیاز دارید، اما اگر فقط می خواهید با نحوه کار سرور آشنا شوید، می توانید حتی عدد ۲ را واردکنید. به این ترتیب آنالیز بسیار ساده خواهد بود، اما انتظار یک جواب با معنی از نظر بیولوژیکی نداشته باشید! به علاوه زمانیکه اعداد بزرگتر را انتخاب می کنید، سرور نتیجه را به آدرس پست الکترونیک شما می فرستد.

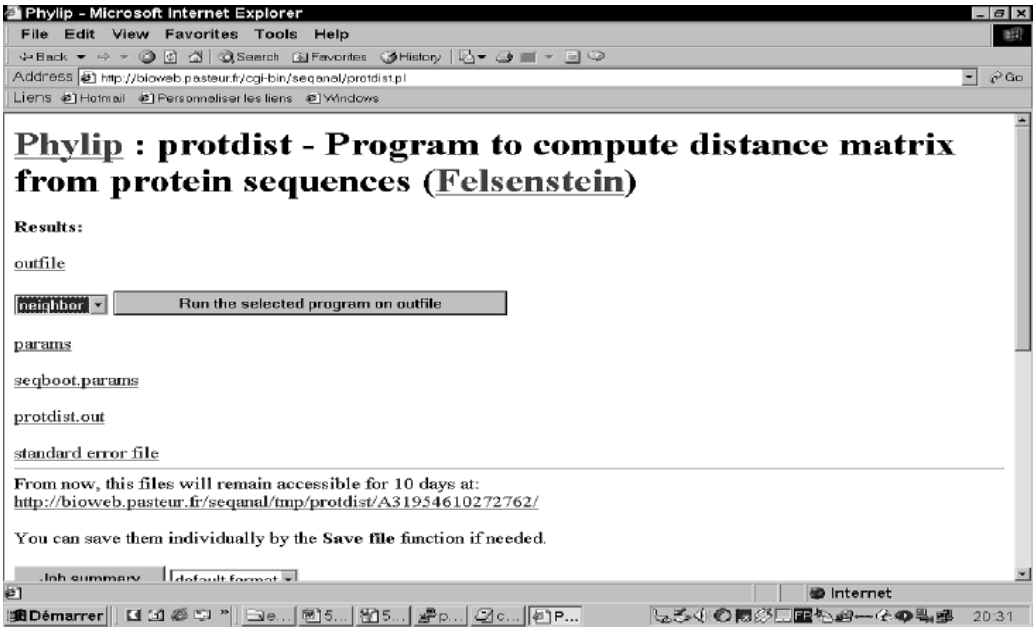
۹. به ابتدای صفحه برگردید و روی دکمه Run Protdist کلیک کنید.

اکنون صفحه نتایج را می توانید مشاهده کنید (شکل ۱۰-۱۳). اگر زمان استفاده مرورگر شما تمام شد (Browser Time Out)، سرور یک نامه الکترونیک حاوی یک فایل URL برای شما می فرستد. اگر این فایل را باز کنید، صفحه نتایج مشابه شکل ۱۰-۱۳ باز می شود.

۱۰. گزینه Neighbor را از اولین منوی موجود در صفحه انتخاب کنید.

Neighbor برنامه ای است که ماتریکس فاصله را به یک درخت Neighbor Joining تبدیل می کند. اگر تمایل داشته باشید، می توانید از روش های دیگر بازسازی درخت (مثل Fitch، که از متد -mean-least square استفاده می کند)، استفاده کنید. امروزه اغلب اعتقاد بر این است که روش NJ نسبت به سایر روش ها بهتر عمل می کند.

[1] number of replicates



شکل ۱۰-۱۳: نتایج Protdist

۱۱. روی دکمه Run the Selected Program on Outfile کلیک کنید.

صفحه اصلی Neighbor ظاهر می شود.

۱۲. در خطی با عنوان Distance Method، نوع درختی را که می خواهید بسازید را مشخص نمایید (شما باید گزینه Neighbor Joining را انتخاب نمایید).

روش های Neighbor Joining و UPGMA، روش های مرتبطی هستند، اما روش Neighbor Joining نتایج بسیار صحیح تری ایجاد می نماید.

روش UPGMA درخت های ریشه دار ایجاد می نماید. منظور از ریشه در این درخت ها، حدس هایی است که در مورد جد مشترک زده می شود (شکل ۱۴-۱۳ را ببینید).

روش Neighbor Joining درخت های بدون ریشه ایجاد می کند، یعنی سعی نمی کند محل جد مشترک را حدس بزند. شما بعداً می توانید خودتان به درخت حاصل ریشه اضافه کنید.

۱۳. لینک Bootstrap options را انتخاب کنید.

با این کار به بخش های پایین تر صفحه می آیید. در این قسمت باید اطلاعاتی راجع به Bootstrapping که انجام داده اید، وارد کنید (به شکل ۹-۱۳ رجوع کنید).

این بخش شبیه به بخشی که در مراحل ۵، ۷ و ۸ دیدید، نیست. در بخش های مذکور، شما تعیین کردید که چه تعداد دوره Bootstrap انجام شود. اکنون در این مرحله، شما باید به خاطر بیاورید که چه تعداد دوره برای برنامه تعیین کرده اید!

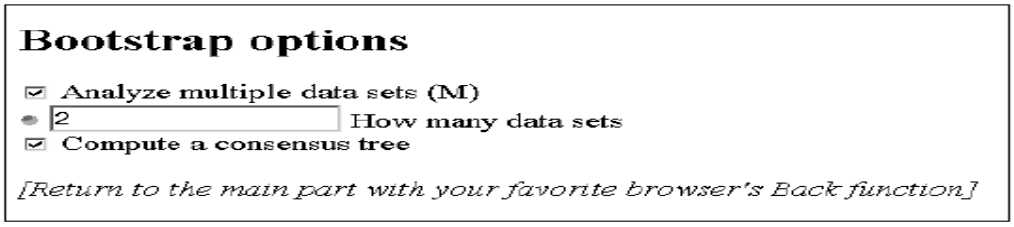
۱۴. گزینه Analyze Multiple Data Sets را انتخاب کنید.

۱۵. تعداد مجموعه های داده هایی را که آماده کرده اید را وارد کنید (شکل ۱۱-۱۳).

هر عددی را که در مرحله ۸ انتخاب کرده اید، در این قسمت وارد کنید.

۱۶. در قسمت Random Seed یک عدد فرد وارد کنید (مثلا ۱ یا ۳)، سپس مربع کنار Randomize را تیک بزنید.

لازم نیست عددی که در این قسمت انتخاب می کنید مشابه عددی باشد که در مرحله ۷ انتخاب کرده اید.



شکل ۱۱-۱۳: بخش Bootstrap options

۱۷. مربع کنار گزینه Compute a Consensus Tree را تیک بزنید.

این گزینه به برنامه فرمان می دهد که درختی را که با مجموعه اطلاعات تعیین شده مربوط به تعداد تکرار ها، بیشترین مطابقت و سازگاری را دارد، به شما ارائه دهد. این گزینه فقط زمانیکه از bootstrap استفاده می کنید مفهوم پیدا می کند، و اعدادی بزرگتر از ۱ در مراحل ۸ و ۱۵ برای آن انتخاب کرده اید که نمایانگر چندین مجموعه داده هاست (Multiple Datasets).

۱۸. به بخش های پایینتر صفحه بیایید تا به قسمت Other Parameters برسید. یک outgroup را انتخاب کنید.

در این قسمت باید به چند نکته توجه داشته باشید:

- انتخاب یک outgroup فقط زمانی معنی پیدا می کند که شما در مرحله ۱۰ روش Neighbor Joining را انتخاب کرده باشید.
- شما می توانید انتخاب کنید که کدام outgroup در کدام محل از توالی های تطابق یافته قرار گیرد، و بدین ترتیب Neighbor از آن outgroup به عنوان ریشه درخت استفاده می کند.
- در درخت شما، ریشه نشاندهنده جد مشترک همه توالی ها است. این outgroup ریشه واقعی نیست، اما ریشه واقعی لزوماً یک چیزی بین outgroup و نزدیک ترین جوانه به آن است.

گاهی ClustalW بسته به سرور و گزینه های انتخاب شده، زمانیکه می خواهد توالی های شما را تطابق دهد، ترتیب آن ها را تغییر می دهد. به هر حال زمانیکه شما یک outgroup را انتخاب می کنید، ترتیب توالی های تطابق یافته رعایت خواهد شد. بنابراین ما پیشنهاد می کنیم که یک نگاهی به تطابق چند توالی خود بیندازید، تا مطمئن شوید که outgroup درست را انتخاب می کنید.

۱۹. به بالای صفحه برگردید و روی دکمه Run Neighbor کلیک کنید.

نتایج شما در یک صفحه جدید به صورت یک مجموعه از لینک ها ظاهر می شود.

۲۰. فایل های مربوط به درخت اجماع (Consensus)را باز کنید.

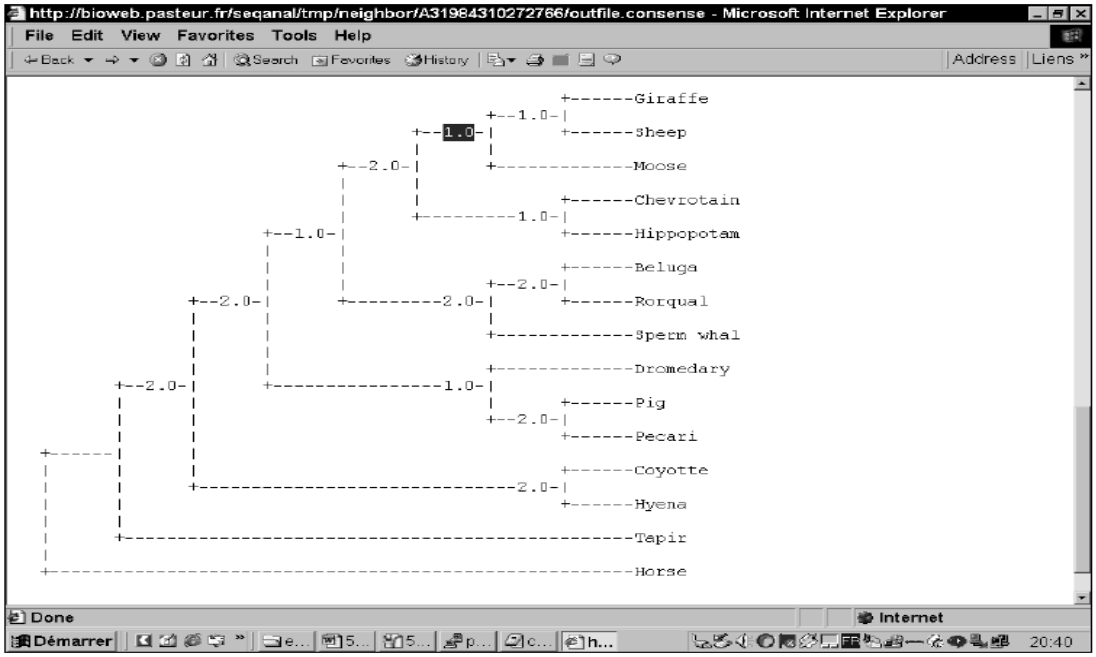
دو فایل ابتدایی در لیست روبروی شما، فایل مربوط به درخت consensus هستند. این فایل ها به صورت لینک هستند؛ بنابراین شما می توانید با کلیک روی آن ها، محتویات آن ها را ببینید. فایل outfile.consense (شکل ۱۲-۱۳) حاوی یک نسخه از درخت فیلوژنتیکی شما با فرمت text است.

درخت حاصل همیشه بدون ریشه است. شماره بالای هر شاخه میزان پایداری آن را نشان می دهد. عدد ۵۰ یعنی پایداری بیشتر و عدد ۲۰ یعنی پایداری کمتر.

فایل outtree.consense همان درخت را در فرمت Newick نشان می دهد.

۲۱. درخت های معمول را باز کنید.

برنامه Neighbor همچنین درخت هایی را که برای ساخت درخت consensus استفاده کرده است، در دو فایل outfile و outtree نشان می دهد.



شکل ۱۲-۱۳: یک درخت consensus ایجاد شده توسط Neighbor

- فرمت outfile حاوی یک نمایش از درخت شما به صورت text است. تعداد درخت ها در این قسمت برابر تعداد تکرار هایی است که در مرحله ۸ و ۱۶ انتخاب کرده اید.
- اگر در مرحله ۱۲ مربع کنار Neighbor Joining را تیک زده باشید، outgroup ریشه درخت شما قرار داده شده است.
- اگر UPGMA را انتخاب کرده باشید، درخت حاصل خود به خود دارای ریشه است.
- فایل outtree حاوی درخت ها به فرمت Newick است.

۲۲. نتیجه را ذخیره کنید و سپس دوباره آن را باز کنید.

نمایش دادن یک درخت با ارزش Bootstrap با ابزار های آنلاین مشکل است. برای حفظ این ارزش ها، بهترین راه اینست که outfile.consense را درون درخت ذخیره نمایید (این یک فایل متن استاندارد است که با استفاده از گزینه Save As در منوی file قابل ذخیره شدن است.).

اگر می خواهید درخت ذخیره شده را با استفاده از یک فرمت گرافیکی دیگر نمایش دهید، راحت ترین راه اینست که ابتدا outtree.consense را ذخیره کنید، سپس آنر توسط سرور Phylodendron به فرمت GIF یا PDF تبدیل کنید (چگونگی انجام این کار در پایان این قسمت توضیح داده شده است).

Bootstrapping چیست؟

اینکه درخت شما به وسیله یک کامپیوتر ایجاد شده است، لزوماً به این معنی نیست که آن درخت کاملاً صحیح است. در مسیر ساخت درخت اشتباهات زیادی ممکن است رخ دهد، مثلا داده های ناقص یا تطابق غیر صحیح توالی ها. در اکثر موارد، درخت حاصل در کل صحیح است، اما حاوی چند شاخه غیر صحیح است. در این موارد Bootstrapping به کار شما می آید. Tree Bootstrapping یک راه برای بررسی اینست که آیا درخت شما از نظر بیولوژیکی با معنی هست یا خیر؟ در

واقع فرایند ارزیابی قدرت[1] درخت شماست، و بررسی می کند که آیا قسمت های مختلف توالی های تطابق یافته به میزان مساوی با درخت شما سازگاری دارند؟ فلسفه Bootstrapping اندازه گیری استواری هر شاخه از درخت و تعیین میزان اعتبار آن برای استفاده در نتیجه گیری ها است.

Bootstrapping یک فرایند چند مرحله ای است. این برنامه با ستون های نمونه گیری در ردیف های اولیه توالی ها آغاز می شود. هدف از این برنامه تولید یک Bootstrap alignment است که در آن بعضی ستون ها حذف و بعضی دیگر دو برابر شده اند. مراحل انجام کار به شرح زیر است:

۱. به صورت تصادفی یک ستون را از توالی های تطابق یافته اولیه انتخاب کنید.

این ستون اولین ستون Bootstrap alignment شما است.

۲. یک ستون دیگر در توالی های ردیف شده اولیه انتخاب کنید، آنرا کپی کنید تا Bootstrap alignment گسترش پیدا کند. گسترش پیدا کردن اینجا یعنی Bootstrap alignment را یک

ستون از ستونی که در مرحه اول از alignment اولیه کپی کرده اید، بلند تر نمایید.

۳. مرحله ۲ را آنقدر تکرار کنیدکه Bootstrap alignment شما به اندازه alignment اولیه ستون داشته باشد. توجه کنید که بعضی ستون های alignment اصلا انتخاب نمی شوند و بعضی دیگر به نظر می رسد که چندین بار در Bootstrap alignment تکرار شده اند. به این فرایند نمونه گیری تصادفی می گویند.

به عنوان مثال، در زیر نشان داده شده که چگونه از یک alignment اولیه می توان دو Bootstrap alignment به وجود آورد:

```
Initial Alignment
Column  1 2 3 4 5 6 7 8 9
seq1    A B C D E F G H I
seq2    A A B B C B A C A
seq3    C C A C B A C A B

Bootstrap Alignment 1
1 1 8 1 2 5 1 8 2
A A H A B E A H B
A A C A B C A C A
C C A C C B C A C

Bootstrap Alignment 2
1 4 5 6 6 3 4 1 7
A D E F F C D A G
A B B C C B B A A
C C B A A A C C C
```

توجه کنیدکه ستون شماره ۱ در Bootstrap alignment 1 چهار بار تکرار شده است.

با این فرایند چندین multiple alignment بوجود می آید که کم و بیش شبیه alignment اولیه هستند. هدف از همه این کارها بررسی اینست که آیا همه ستون ها در alignment اولیه شما به یک فرایند تکاملی مشابه اشاره دارند یا خیر؟

در مرحله بعد هر alignment تصادفی به یک ماتریکس فاصله، و در نهایت هر ماتریکس به یک درخت تبدیل می شود. برنامه Phylip برای ساخت درخت consensus، از همه درخت های ساخته شده به این روش میانگین می گیرد.

برنامه Phylip برای ارزیابی کیفیت هر شاخه در درخت consensus، تعداد درخت های bootstrap را که حاوی آن شاخه هستند شمارش می کند. شاخه های خوب آن هایی هستند که در همه درخت های bootstrap ظاهر می شوند.

[1] robustness

به شکل ۱۲-۱۳ برگردید و نگاهی به درخت consensus بیندازید. در این درخت consensus همراه هر شاخه یک عدد یا ارزش بین ۱ و ۲ آمده است. این عدد میزان استحکام و اعتبار آن شاخه را نشان می دهد. این عدد به این دلیل بین ۱ و ۲ است که در مراحل ۸ و ۱۵ تعداد دوره ها را ۲ انتخاب کردیم. Neighbor دو درخت bootstrap با استفاده از این دو bootstrap alignment ایجاد کرده است، و درخت consensus نهایی میانگینی از آن دو درخت است.

در یک درخت، یک شاخه همیشه داده های شما را به دو بخش تقسیم می کند: توالی های سمت چپ و توالی های سمت راست شاخه. اعداد بالای هر شاخه در درخت consensus نمایانگر این هستند که چه تعدادی شاخه در دو درخت شما وجود دارند که داده ها را با روشی دقیقاً مشابه با این شاخه، به دو بخش تقسیم می کنند.

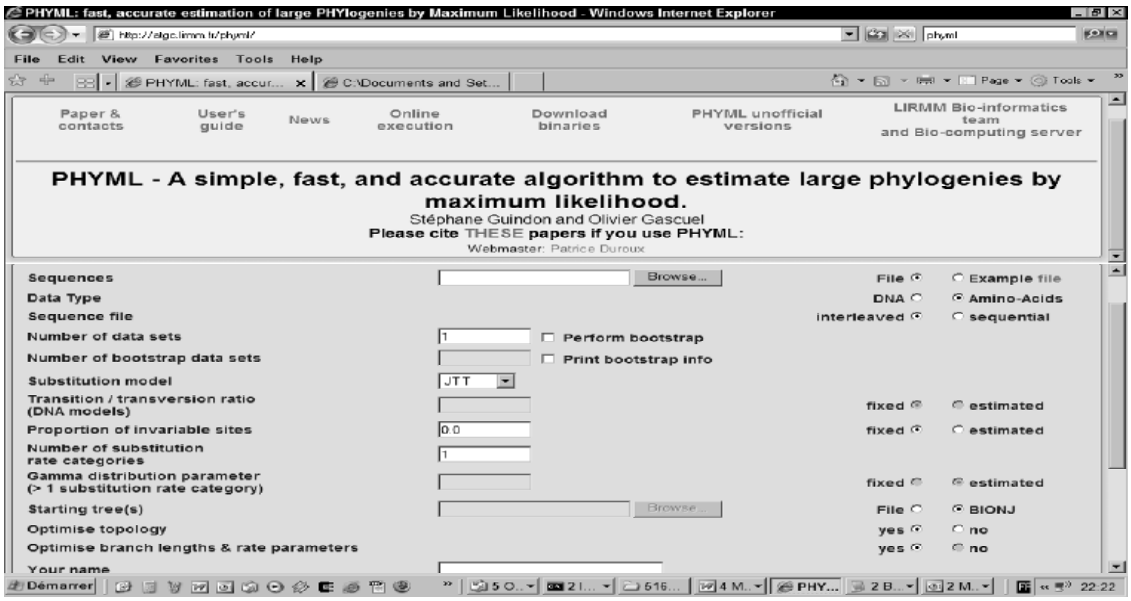
به عنوان مثال، اگر به درخت consensus در شکل ۱۲-۱۳ نگاه کنید، متوجه می شوید که شاخه های حاوی زرافه و گوسفند دارای ارزش ۱ هستند. دلیل آن اینست که در یکی از درخت های bootstrap، زرافه و گوزن شمالی در یک گروه هستند، در حالیکه در درخت دیگر زرافه و گوسفند در یک گروه قرار دارند (و گوزن شمالی به تنهایی در یک گروه قرار گرفته است). از دیدگاه تئوری، می توان نتیجه گرفت که در مورد اینکه آیا گوسفند و زرافه خویشاوندی نزدیکتری دارند یا هر کدام با گوزن شمالی، عدم اطمینان وجود دارد. در عمل، فقط با دو درخت bootstrap، نمی توان نتیجه گیری انجام داد. برای یک نتیجه گیری مطمئن به حداقل ۱۰۰ دوره bootstrap نیاز است.

ساختن درخت بیشترین احتمال[1] با استفاده از PhyML

به نظر می رسد که درخت های بیشترین احتمال صحیح تر از سایر درخت ها به نظر می رسند، چون این ها درخت هایی هستند که با بیشترین احتمال (از نظر آماری)، alignment شما را تفسیر می کنند. به عبارت دیگر alignment شما یک داستان کوتاه است در مورد اینکه چگونه یک مجموعه از جهش ها، که از توالی جد مشترک شروع می شوند، منجر به ایجاد تنوع موجودات می شوند. روش بیشترین احتمال تضمینی است برای اینکه درخت و alignment دقیقاً به یک روند اشاره دارند. پیش از این، اطلاعات مربوط به درخت های بیشترین احتمال، به مجموعه های کوچکی از داده ها محدود بودند، چون طراحی آن ها بسیار وقت گیر بود. اما کشف جدید Olivier Gascuel و Stephane Guindon (دانشگاه Montpelier، فرانسه)، باعث شده که امروزه طراحی اینگونه درخت ها در زمان کوتاه تری صورت گیرد. نحوه استفاده از PhyML در زیر آمده است:

۱. در آدرس اینترنتی www.tcoffee.org/dummy_aln3.html ، alignment تقلبی را باز کنید.
این alignment با فرمت Phylip است، یعنی تنها فرمتی که توسط این سرور شناخته می شود.
۲. این alignment را با استفاده از گزینه Save As در منوی File، با فرمت text ذخیره نمایید.
۳. به آدرس اینترنتی atgc.lirmm.fr/phyml/ وارد شوید.
۴. دکمه رادیوی File را انتخاب کنید (شکل ۱۳-۱۳).
اکنون می توانید با استفاده از دکمه Browse، فایل خود را به این قسمت منتقل نمایید.
۵. دکمه رادیویی Amino Acids را انتخاب کنید.
alignment ما حاوی آمینو اسید است.

[1] maximum likelihood tree



شکل ۱۳-۱۳: آغاز کار با سرور PHUML

۶. دکمه رادیویی Interleaved را انتخاب کنید.

اکنون alignment ما به فرم Phyllips Interleaved است، یعنی به قطعاتی تقسیم شده است (برای اطلاع از فرمت های مختلف به فصل ۹ مراجعه کنید.)

۷. مربع کنار گزینه Perform Bootstrap را تیک بزنید، سپس در قسمت Number of Bootstrap Data Sets عدد ۱۰۰ را وارد کنید.

انجام یک bootstrap زمان زیادی لازم دارد، اما برای اینکه میزان معنی داری درخت خود را از نظر بیولوژیکی تخمین بزنید، یک مرحله ضروری است.

در عمل بهتر است ۱۰۰۰ دوره bootstrap انجام دهید، اما برای یک مثال ۱۰۰ دوره کافی است.

۸. در پایین صفحه به قست های fill in the Name و E-Mail Address بروید.

۹. روی دکمه Run کلیک کنید.

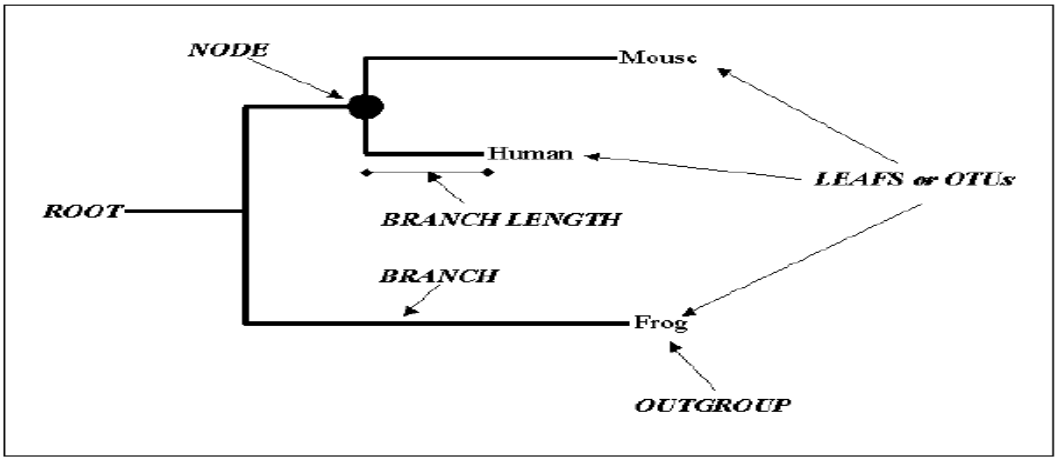
بعد از انجام این مراحل، حدود نیم تا ۱ ساعت طول می کشد تا نتایج به پست الکترونیک شما فرستاده شوند. این نتایج شامل یک درخت به فرمت Newick است که می توانید آنرا با استفاده از Phylodendron به فرمت های دلخواه خود تبدیل کنید (این موضوع در ادامه توضیح داده خواهد شد). این درخت حاوی اعداد bootstrap است، که میزان اعتبار هر جوانه[1] را نشان می دهند. هر چه این عدد بزرگتر باشد، آن جوانه قابل اعتماد تر است.

اجزاء متشکله یک درخت

زمانیکه با درخت های فیلوژنتیکی کار می کنید، به یک واژه نامه خاص هم نیاز دارید. شکل ۱۴-۱۳ نامهای بخش های مختلف یک درخت را نشان می دهد. بعضی از این نام ها ممکن است کمی تکنیکی به نظر برسند، اما بهتر است معنی آن ها را بدانید، چون اصطلاحات فیلوژنتیکی هر روز بیش از پیش در انتشارات معمول می شوند.

[1] node

برگ ها با عنوان OTUs یا واحد های تاکسونومیک عملکردی[1] هم شناخته می شوند. هر جوانه نشاندهنده یک OUT اجدادی است. clade به یک گروه از OUT ها گفته می شود که حاوی چندین توالی و جوانه های مربوط به جد مشترکشان هستند. ریشه، جد مشترک همه OUT ها است.



شکل ۱۴-۱۳: بخش های مختلف یک درخت فیلوژنتیکی

شاخه، رابطه بین یک clade یا یک OUT با بقیه درخت را بیان می کند. یک درخت می تواند مقیاس دار[2] یا بدون مقیاس[3] باشد. در درخت مقیاس دار طول شاخه به معنای زمان تکامل است، اما در درخت بدون مقیاس فقط محل شاخه ها مد نظر و معنی دار است.

همچنین درخت می تواند ریشه دار یا بدون ریشه باشد. درخت های بدون ریشه مسیر های تکاملی را مشخص نمی کنند. اگر درخت شما بدون ریشه است، باید این احتمال را بپذیرید که هر یک از جوانه ها یا OUT ها می توانند نسبت به OUT های دیگر به جد مشترک نزدیک تر باشند. اگر درخت حاوی outgroup باشد، ریشه باید بین این outgroup و دیگر clade ها قرار گیرد.

ریشه یک درخت قدیمی ترین جد مشترک در آن درخت است. این مطلب خوب به نظر می رسد، اما متأسفانه در مورد درخت هایی که ما در اینجا می سازیم، بی معنی است. مشکل آن ها اینست که روش های بازسازی درخت جهت تکامل را مشخص نمی کنند. با این روش های نمی توان تعیین کرد که کدام جوانه قدیمی ترین آن ها است. در نتیجه ریشه ای که در بعضی درخت ها مشاهده می کنید، معمولاً قراردادی است، و بنابراین از نظر بیولوژیکی معنی دار نیست. زیست شناسان ترجیح می دهند از درخت های بدون ریشه استفاده کنند تا باعث سردرگمی آنها نشود. اما با این حال اگر باز هم دوست دارید برای درختتان ریشه قرار دهید، تنها راه مناسب اینست که از outgroup استفاده کنید. outgroup ارگانیسمی است که ارتباط آن با سایر ارگانیسم های درخت آنقدر دور است که می تواند با اطمینان به عنوان ریشه مطرح شود. به عنوان مثال، اگر درخت شما حاوی پریمات ها[4] باشد، می توانید از موش به عنوان outgroup استفاده کنید. همچنین اگر درخت شما برای مقایسه پستانداران باشد، می توانید یک خزنده را به عنوان outgroup قرار دهید.

[1] Operational Taxonomic Units
[2] scaled
[3] unscaled
[4] primates

نمایش درخت فیلوژنتیک

Phylodendron یک سرور قدرتمند است که امکان تبدیل یک فایل متن به فرمت Newick را به نمایش های گرافیک برای شما فراهم می کند. با این برنامه شما می توانید کوچک ترین جزئیات نحوه نمایش درخت را کنترل کنید و تصویر هایی با هر فرمت دلخواه ایجاد نمایید. در زیر مراحل تبدیل یک درخت فیلوژنتیک به یک تصویر GIF را، که در هر متنی قابل باز گذاری است، به شما نشان می دهیم:

۱. به آدرس اینترنتی iubio.bio.indiana.edu/treeapp/treeprint-form.html وارد شوید. به عنوان یک مسیر دیگر می توانید از Spanish EMBnet به آدرس www.es.embnet.org/Doc/phylodendron /treeprint-form.html هم استفاده کنید.
۲. درخت خود را با فرمت Newick در پنجره Input وارد کنید.
۳. در بالاترین ردیف در زیر قسمت Tree Styles، دکمه رادیویی Phenogram را انتخاب کنید.
اگر گزینه Tree Diag را انتخاب کرده اید، باید حتماً یک درخت بدون ریشه را به برنامه ارائه کنید، در غیر اینصورت برنامه متوقف می شود.
۴. در قسمت Extra Options، گزینه GIF image را از منوی Format انتخاب کنید.
۵. اگر می خواهید طول شاخه ها معنی دار باشد (درخت مقیاس دار)، مربع کنار گزینه Use Node Lengths در قسمت Tree Growth را تیک بزنید.
۶. روی دکمه Submit کلیک کنید.
۷. با استفاده از گزینه Save As در منوی File، نتایج خود را با فرمت GIF ذخیره کنید.

انجام فیلوژنی به صورت رایگان در اینترنت

انجام فیلوژنی در اینترنت هنوز مانند جستجو در بانک ها اطلاعاتی یا روش های آنالیز توالی ها پیشرفت نکرده است. اما خوشبختانه در چند سال اخیر، چند منبع با کیفیت در دسترس قرار گرفته اند. هر چند این منابع ممکن است سرور های قدرتمندی نداشته باشند، اما اگر هدف شما فقط اینست که گاه گاه یک درخت بسازید، این منابع کارامد هستند.

یافتن منابع آنلاین

تعداد بسیار زیادی منابع فیلوژنی در اینترنت در دسترس هستند. اما بسیاری از آنها قابلیت استفاده آنلاین را ندارند و باید آنها را روی کامپیوتر خود نصب کنید. جدول ۲-۱۳ چند منبع را که می توانید به صورت آنلاین استفاده کنید، لیست کرده است.

Table 13-2 Online Sites for Making Phylogenetic Trees	
Address	***Description***
www.ebi.ac.uk/clustalw/	You can use ClustalW to build multiple alignments and compute NJ trees. Remember: You cannot do both at the same time!
www.genebee.msu.ru/clustal/basic.html	The Genebee server can produce genuine phylogenetic trees in one step.
www.tcoffee.org	Tcoffee computes a genuine NJ-phylogenetic tree in one step.
www.jalview.org	You can use Jalview to produce NJ trees. It's a very powerful tool that combines alignment editing with tree computation.
atgc.lirmm.fr/phyml/	A powerful method to compute maximum likelihood trees from Gascuel and his team.
bioweb.pasteur.fr/seqanal/interfaces/bionj-simple.html	An interface to BioNJ, a novel NJ method.
www.up.univ-mrs.fr/evol/figenix/	A powerful Java tool to gather members of a protein family and build the associated tree.
bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html	A Web interface for Phylip.
www.genebee.msu.ru/services/phtree_reduced.html	Very powerful interface for a novel tree-reconstruction method.

یافتن منابع عمومی

جدول ۳-۱۳ بعضی از مهمترین منابع فیلوژنی را لیست کرده است. اغلب شما می توانید هر چیزی را که نیاز دارید فقط با چند کلیک در آدرس های زیر پیدا کنید. در این قسمت همچنین دو منبع که پایه های فیلوژنی را توضیح می دهند، ارائه شده اند.

Table 13-3 Generic Phylogenetic Resources on the Internet	
Address	***Description***
evolution.genetics.washington.edu/phylip/software.html	Joe Felsenstein's pages, where Phylip lives; it's also one of the most extensive collections of resources available. Truly a legendary site!
www.ucmp.berkeley.edu/subway/phylo/phylosoft.html	A very complete list of phylogeny resources.
paup.csit.fsu.edu/index.html	The home of PAUP, legendary phylogeny package using Parsimony. Although PAUP is a commercial package, it's reasonably priced and worth every penny, according to specialists.
www.ncbi.nlm.nih.gov/About/primer/phylo.html	The NCBI primer on phylogeny.
www.techfak.uni-bielefeld.de/bcd/Curric/MathAn/mathan.html	A high-quality course on tree reconstruction methods.

مجموعه های ژنهای اورتولوگ

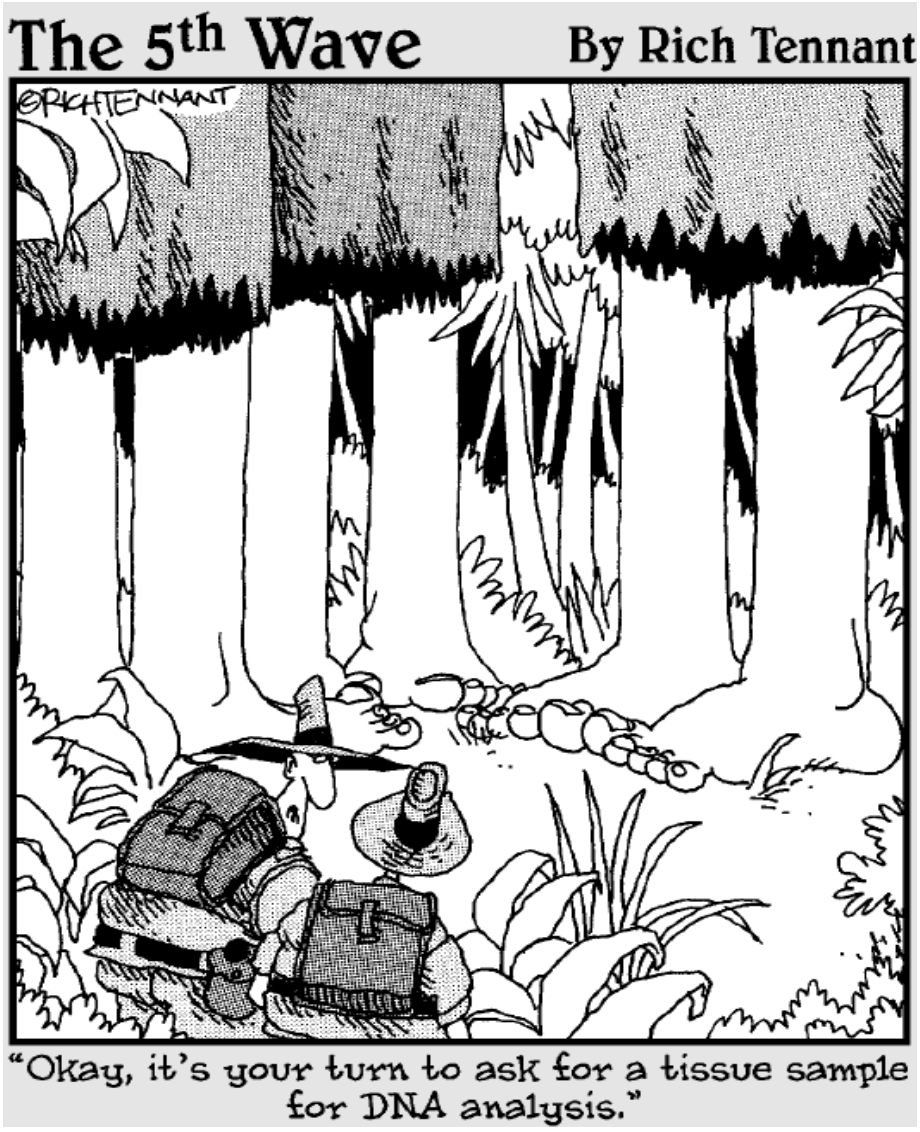
زمانیکه سؤالاتی در مورد عملکرد و فیلوژنی داشته باشید، مجموعه هـای ژن هـای اورتولـوگ و همولـوگ، بـسیار مفیـد هستند. جدول ۴-۱۳ بعضی بانک های اطلاعاتی که این اطلاعات را ارائه می دهند، لیست شده اند.

Table 13-4 Collections of Orthologous Sequences

Address	*Description*
`www.ncbi.nlm.nih.gov/COG/`	Cluster of orthologous sequences maintained by the NCBI. Each cluster contains proteins from bacterial genomes.
`pbil.univ-lyon1.fr/databases/hovergen.html`	A collection of orthologous vertebrate genes.
`pbil.univ-lyon1.fr/databases/hogenom.html`	A collection of orthologous bacterial genes.
`systers.molgen.mpg.de`	Another collection of homologous sequences.
`rdp.cme.msu.edu/`, `www.psb.ugent.be/rRNA/ssu/`, `www.psb.ugent.be/rRNA/lsu/`	Three extensive collections of ribosomal RNA sequences, which are very useful for classifying new organisms and come with appropriate phylogenetic tools.

بخش پنجم

قسمت دهگانه



ترجمه: سجاد عباسی

ویرایش: دکتر حسن کربکندی

در این بخش می خوانیم...

این بخش حاوی جدول هایی از لینک ها، منابع، و مطالب مفید دیگر است. اگر هنوز مطلبی را که در جستجوی آن هستید در این کتاب پیدا نکرده اید، ممکن است آنرا در این بخش، یا در یکی از لینک هایی که ارائه شده پیدا کنید.

اگر می خواهید تحقیقات خود را بر پایه اینترنت انجام دهید، در این بخش چند رهنمود مفید به شما ارائه می کنیم. بهتر است قبل از اینکه خیلی دیر شود، این بخش را بخوانید!

فصل چهاردهم

ده (خیلی خوب! دوازده) دستورالعمل برای استفاده از سرور ها

در این فصل می خوانیم:

- ✓ اطمینان از اینکه داده های سری را فاش نکرده اید
- ✓ اطمینان از اینکه می توانید کار هایی را که تاکنون انجام داده اید، تکرار کنید
- ✓ استفاده از داده های مناسب
- ✓ اطمینان از اینکه فایل های درست را ذخیره کرده اید

.....

"من هرگز به کسی دروغ نگفتم. من دقیقاً حقیقت را گفتم، اما آن ها فکر کردند که دروغ است."

— Harry S. Truman (1884–1972)

جالب ترین چیز در مورد استفاده از سرور های اینترنتی از دیدگاه کاربران اینست که در زمینه بیوانفورماتیک هیچ مشکلی در اینترنت نخواهید داشت. تصور کنید! شما می توانید از هر برنامه ای که بخواهید، استفاده کنید. شما نیازی ندارید که هیچ بانک اطلاعاتی را نگهداری یا به روز رسانی کنید.

اینجا همه چیز خوب است، اما زمانیکه با استفاده از این ابزار های بیوانفورماتیک بخواهید جستجو انجام دهید، باید کمی بیشتر دقت کنید. هیچ چیز نا امید کننده تر از این نیست که شما نتوانید یک نتیجه خوب را دوباره تکرار کنید، بخصوص در زمینه علوم.

در این فصل چند توصیه برای راحت تر کردن کار در زمینه بیوانفورماتیک، به شما ارائه می کنیم.

به خاطر بسپارید: داده های شما بر روی وب هرگز در امان نیستند!

زمانیکه یک توالی را به یک سرور در اینترنت ارائه می دهید، در واقع آنرا وارد دنیای بی انتها کرده اید. به استثنای چند برنامه کاربردی Java که به صورت offline هم قابل استفاده هستند، هرگز داده های خود را با اطمینان در اینترنت ارائه نکنید.

با وجود اینکه مسئولین این سرور ها به درستکاری مشهور هستند، اما ایمنی داده های شما را به هیچ وجه تضمین نمی کنند. یعنی داده های شما ممکن است در راه انتقال به این سرور ها دزدیده شوند. به عقیده ما هیچ کس از کنار داده هایی که شما وارد اینترنت می کنید، به سادگی عبور نمی کند.

اگر می خواهید داده های شما مخفی باقی بماند، باید برنامه های لازم برای کار خود را روی کامپیوتر شخصی تان نصب کنید. در این صورت جای هیچ اما و اگری باقی نمی ماند.

سرور، بانک اطلاعاتی، و نسخه ی از برنامه های مورد استفاده را به خاطر بسپارید

اگر در دو سرور مختلف، با یک برنامه (مثلاً ClustalW) کار می کنید، اصلاً بدین معنی نیست که دو کار مشابه را دو بار انجام می دهید. امکانات هر سرور و حتی نسخه برنامه های آن ها ممکن است متفاوت باشد. به همین خاطر جمله " من در اینترنت با برنامه ClustalW کار می کنم"، حاوی اطلاعات کافی برای بازیابی نتایج مورد نظر توسط یک فرد دیگر نیست.

برای جلوگیری از ایجاد اشتباه، نام سرور و همچنین نام نسخه ای از هر برنامه که استفاده می کنید را بنویسید. به عنوان مثال نمی توان انتظار داشت که سروری که با ClustalW 1.77 کار می کند با سروری که با ClustalW 1.81 کار می کند، نتایج مشابهی داشته باشند.

اگر برنامه شما، مثلاً (BLAST)، از یک بانک اطلاعاتی (مثلاً Swiss-Prot) استفاده می کند، شما باید نام و نسخه بانک اطلاعاتی که جستجو می کنید را بنویسید.

سرور های همچنین در طول زمان تغییر می کنند. به طور متوسط این اتفاق هر ۶ ماه یک بار می افتد.

شماره شناسایی توالی ها را یادداشت کنید

زمانیکه یک توالی را آنالیز می کنید، مطمئن شوید که اطلاعات کافی در مورد آن توالی دارید. سعی کنید شماره شناسایی و شماره دسترسی آن، و همچنین هر گونه اطلاعات مربوط به آن را نگهداری کنید. توجه کنید که شماره شناسایی یک ورودی در بانک های اطلاعاتی مختلف ممکن است متفاوت باشد، اما شماره دسترسی همیشه یکسان است.

ویژگی های برنامه را یادداشت کنید

زمانیکه از یک سرور استفاده می کنید، اغلب لازم است که بعضی از گزینه ها را تغییر دهید. بعضی برنامه ها همراه با نتایج، یک فایل حاوی تمام پارامترهایی که شما انتخاب کرده اید به شما ارائه می دهند. این فایل اغلب یک نام دارد که حاوی کلمه log است. در مورد BLAST این پارامتر ها در پایین صفحه نتایج نشان داده می شوند. اگر چنین فایلی همراه نتایج دیدید، حتماً آنرا ذخیره کنید.

به عنوان یک راه فرعی، می توانید با استفاده از ابزار PrntScrn ویندوز از صفحه مورد نظر عکس بگیرید (نکاتی در این مورد در دستور العمل بعدی ذکر شده است).

نتایج خود را به طور صحیح ذخیره کنید

ذخیره کردن نتایج اینترنتی در نگاه اول ساده به نظر می رسد، اما نکاتی در آن نهفته است. این مطلب بخصوص زمانیکه می خواهید یک سند حاوی عکس، مثل نتایج BLAST، را ذخیره کنید، مهم است. برای این کار نمی توان یک دستورالعمل کلی تجویز کرد، چون بهترین روش برای ذخیره کردن اطلاعات، بسته به جستجوگر وب، چگونگی نصب جستجوگر روی کامپیوتر شما، و نوع اطلاعاتی که می خواهید ذخیره کنید، همیشه در حال تغییر است. هنگام ذخیره کردن فایل ها، Java منبع اصلی دردسر است، و معمولاً تنها راه حل در این مواقع، گرفتن عکس از کل صفحه است.

در مورد اطلاعات مهم به یک بار ذخیره کردن اکتفا نکنید، همیشه دو بار آنرا ذخیره کنید. آخرین نسخه از Internet Explorer و Netscape امکان ذخیره کل صفحه (از جمله عکس ها) را فراهم می کنند.

اگر نتوانستید صفحه مورد نظر را ذخیره کنید، سعی کنید آنرا به صورت PDF پرینت کنید. اگر موفق نشدید این کار را هم انجام بدهید، با استفاده از ابزار Windows PrntScrn یک عکس از صفحه بگیرید:

۱. دکمه PrntScrn صفحه کی بورد را فشار دهید.

با این کار یک کپی از کل صفحه در Clipboard ذخیره می شود.

۲. برنامه Microsoft Paint را از مسیر زیر باز کنید:

Start⇨Programs⇨Accessories⇨Paint

۳. دکمه های Ctrl + V را فشار دهید، یک پیام ظاهر می شود، در این پیام گزینه Yes را انتخاب کنید.

۴. فایل را به صورت معمولی یا print ذخیره کنید.

فایل را با فرمت jpg. ذخیره کنید تا یک عکس فشرده ایجاد شود.

هر جا ممکن بود یک نسخه text هم از نتایج خود ذخیره کنید. بسیاری از برنامه ها نتایج را در یک نسخه text و نیز یک نسخه گرافیکی زیبا ارائه می دهند. استفاده از نسخه text یک راه بسیار ساده تر برای ذخیره و استخراج نتایج است. در بسیاری از موارد، سایر برنامه ها فقط فرمت text را می توانند بخوانند.

استفاده از E-value

E-value به شما می گوید نتایجی که شما به دست آورده اید چند بار می تواند به صورت شانسی تکرار شود. این عدد به خودی خود از نظر بیولوژیکی معنی ندارد. اما دلیل استفاده از این عدد در زیست شناسی اینست که یک نتیجه خوب بصورت شانسی با احتمال خیلی کمی رخ می دهد.

E-value هر چه کوچکتر باشد بهتر است. به عنوان مثال ارزش E برابر $۱۰^{-۳۲}$ بهتر از $۱۰^{-۴}$ است.

آیا alignment شما قابل اطمینان است؟

صرف اینکه شما یک alignment دارید، دلیل بر این نیست که آن alignment حتماً قابل اطمینان است. قبل از اینکه با استفاده از alignment خود کاری انجام بدهید، مطمئن شوید که از نظر بیولوژیکی صحیح است. درصد یکسانی[1] را اندازه بگیرید و دنبال قطعات محافظت شده[2] یا موقعیت های محافظت شده مهم بگردید. دید نقادانه داشته باشید و البته از رهنمود های ذکر شده در فصل ۹ هم استفاده کنید.

نتایج سر مرز را با استفاده از برنامه های مختلف چک کنید

اگر از نتایج بدست آمده مطمئن نیستید، روش های دیگر را هم امتحان کنید. اگر باز هم همان نتایج بدست آمد، می تواند به این معنی باشد که نتایج شما صحیح و معنی دار هستند. اما اگر نتایج مشابهی به دست نیاوردید، می تواند به این معنی باشد که نتایج شما یک مجموعه بی ارزش است. به عنوان مثال، اگر یک درخت فیلوژنتیک را با دو روش درست کردید، و دو درخت حاصل خیلی با هم متفاوت بودند، با احتمال زیاد هیچ کدام این آن درخت ها، درخت صحیح نیست. زیست شناسان قابلیت تکرار نتایج با استفاده از روش های مختلف را *قدرت*[3] نتایج می نامند، و این مفهوم در مورد اکثر مطالبی که در این کتاب گفته شد، کاربرد دارد (به عنوان مثال بخش "شناسایی دامین های عبور کننده از غشاء" در فصل ۶ را ببینید.).

زمانیکه شما از برنامه های مختلف استفاده می کنید، حتماً از یک برنامه که مبنای کار آن متفاوت است هم استفاده کنید. مثلاً مقایسه یک درخت NJ از ClustalW با یک درخت NJ دیگر از Phylip منطقی نیست. بهتر است درخت NJ را با یک درخت Fitsch یا ML مقایسه کنید.

به عنوان یک راه فرعی، می توانید چند پارامتر اصلی برنامه را تغییر دهید و تأثیر آن را بر درخت حاصل بررسی کنید. مثلاً در یک multiple alignment، شما می توانید گزینه gap penalties را تغییر دهید.

---

[1] percentage of identity
[2] conserved blocks
[3] robustness

از روش های ثبت نشده اجتناب کنید

منظور ما این نیست که همواره از روش های جدید اجتناب کنید، اما استفاده از روش های شناخته شده بهترین راه برای بدست آوردن نتایج قابل اطمینان است. هرگز از یک برنامه شناخته نشده و ثبت نشده که مبنای کار آن را نمی دانید، استفاده نکنید.

بانک های اطلاعاتی با گذر زمان اعتبار خود را از دست می دهند

بانک های اطلاعاتی مانند شراب نیستند که هر چه بمانند بهتر شوند! اطلاعات قدیمی استخراج شده از بانک های اطلاعاتی را استفاده نکنید. هر وقت نیاز به استفاده از یک بانک اطلاعاتی را اطلاعاتی در مورد توالی ها پیدا کردید، مطمئن شوید که حتماً از داده های به روز شده استفاده کنید.

اطلاعاتی که رایگان به نظر می رسند، لزوماً همیشه رایگان نیستند

اطلاعات عمومی همیشه برای هر گونه استفاده ای در دسترس نیستند. قبل از اینکه داده ای را به یک سرور ارائه کنید، مطمئن شوید که اجازه استفاده از آن داده را دارید. مثلاً اگر برای یک شرکت کار می کنید، ممکن است لازم باشد آن داده ها را بخرید.

بسیاری از شرکت ها از ارائه رایگان بعضی داده ها یا سرویس ها، به عنوان تبلیغ استفاده می کنند. این ارائه رایگان کاملاً به مدیر شرکت بستگی دارد. اگر از این منابع استفاده می کنید، ممکن است یک زمانی متوجه شوید که دیگر این شرکت اطلاعات را به صورت رایگان ارائه نمی کند. منظور ما این نیست که هرگز از منابع خصوصی استفاده نکنید، اما همواره باید بدانیدکه این ارائه رایگان

تا چه موقع ادامه دارد، تا خللی در کار شما ایجاد نشود.

چه زمانی برنامه ها را روی کامپیوتر خود نصب کنیم؟

دلیل اینکه این کتاب به منابع آنلاین پرداخته است، اینست که استفاده از این منابع آسان است و مراحل پیچیده راه اندازی را ندارد. با وجود این منابع شما می توانید بجای کامپیوترا، کاملاً روی بیولوژی تمرکز کنید. اگر فقط گاهگاه از بیوانفورماتیک استفاده می کنید، این منابع دقیقاً همان چیزی هستندکه شما نیاز دارید.

اگر متوجه شده اید که شدیداً به بعضی سرور ها اعتماد پیدا کرده اید، برنامه های آن ها را روی کامپیوتر خود نصب کنید. نصب و نگهداری این برنامه های خیلی ساده است. این نرم افزار ها با سیستم های Linux و Unix بهتر کار می کنند. اگر تاکنون با این سیستم ها کار نکرده اید، یاد گرفتن همه اطلاعات لازم برای نصب و کار کردن با Linux احتمالاً چند هفته طول می کشد. اما حتی اگر وقت شما محدود است، صرف کردن این زمان ارزش دارد.

اما از کجا متوجه شویم که زمان آن رسیده که نرم افزار ها را روی کامپیوتر خود نصب کنیم؟ اگر با استفاده از منابع آنلاین ۲۰ بار توانسته اید با BLAST کار کنید، می توان گفت که زمان آن رسیده است.

فصل پانزدهم

چند منبع مفید در بیوانفورماتیک

در این فصل می خوانیم:

- ✓ ده بانک اطلاعاتی مهم
- ✓ چند برنامه خیلی مهم
- ✓ مجموعه های لینک ها و منابع عمومی
- ✓ مطالبی در مورد جدیدترین تکنیک ها

.....

"از کجا فهمیدید که من می دانستم که شما می دانستید؟"

- مثال رایج برنامه ریزی بر مبنای دانش

یکی از خصوصیات منحصر به فرد بیوانفورماتیک اینست که تقریباً هر چیزی که در این زمینه نیاز داشته باشید، بصورت آنلاین دارد. در این فصل منابع اصلی را به شما نشان می دهیم. اگر اطلاعات نسبتاً خوبی در مورد لینک هایی که در این فصل به شما معرفی می کنیم دارید، می توان گفت که در زمینه بیوانفورماتیک نسبتاً خوب آموزش دیده اید.

به خاطر داشته باشید که بیوانفورماتیک به سرعت در حال پیشرفت است. در بیوانفورماتیک مسئله اصلی دانستن مطالب فراوان نیست. شناختن محل هایی که می توانید از آنها اطلاعات مورد نظر خود را استخراج کنید، بسیار مهم تر است. این کتاب همه جزئیات را پوشش نداده است، اما مسیر های کسب اطلاعات را به شما نشان داده است. این مسیر ها را امتحان کنید و مطالب جدید را یاد بگیرید. در این مسیر ها، چیز هایی را که می بینید در گوشه ای از ذهن خود نگه دارید تا زمانیکه به آن ها احتیاج پیدا کنید.

ده بانک اطلاعاتی اصلی

کار بیوانفورماتیک کشف اطلاعات بیولوژیکی است. این اطلاعات در بانک های اطلاعاتی ذخیره شده اند. بانک های اطلاعاتی در واقع موزه های مدرن یا مجموعه های منابع اینترنتی هستند، که همه اطلاعات موجود در یک زمینه را

به دقت ذخیره و دسته بندی کرده اند. در این قسمت زیاد وارد جزئیات نشده ایم، و فقط ۱۰ بانک اطلاعاتی اصلی را معرفی کرده ایم.

Table 15-1 Ten Important Bioinformatics Databases

Name	*Address*	*Has Data About*
GenBank/DDBJ/EMBL	`www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide`	Nucleotide sequences
Ensembl	`www.ensembl.org`	Human/mouse genome
PubMed	`www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed`	Literature references
NR	`www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein`	Non-redundant protein sequences
Swiss-Prot	`www.expasy.ch`	Protein sequences
InterPro	`www.ebi.ac.uk`	Protein domains
OMIM	`www.ncbi.nlm.nih.gov`	Genetic diseases
Enzymes	`www.chem.qmul.ac.uk`	Enzymes
PDB	`www.rcsb.org/pdb/`	Protein structures
KEGG	`www.genome.ad.jp`	Metabolic pathways

ده برنامه نرم افزاری اصلی در بیوانفورماتیک

داده ها به تنهایی ارزشمند نیستند، برای ارزش دادن به آن ها باید نحوه آنالیز آن ها را بلد باشید. برنامه های نرم افزاری بیوانفورماتیک به همین منظور ساخته شده اند. با استفاده از این برنامه ها می توانید داده های خود را جستجو کنید، آن ها را آنالیز کنید و نمایش بدهید. جدول ۲-۱۵ پنج دسته اصلی از برنامه هایی که ممکن است در بیوانفورماتیک به درد شما بخورد را لیست کرده است. همچنین در هر دسته شاخص ترین نرم افزار ها (از نظر خودمان) را معرفی کرده ایم. هر کدام از این برنامه ها در یک بخش از این کتاب معرفی شده اند، که در جدول هم به آن اشاره شده است.

جدول ۲-۱۵ همچنین آدرس یک سرور آنلاین را برای هر برنامه معرفی کرده است. در اکثر موارد، شما می توانید آدرس های دیگر مربوط به هر برنامه را در فصل مربوط به آن برنامه پیدا کنید.

ده یابنده منبع اصلی

بیوانفورماتیک به سرعت در حال تغییر است. جدید ترین برنامه های هم قدیمی و از کار افتاده می شوند. مؤسسه ها تغییر نام می دهند و برنامه های جدیدتر عرضه می شوند. این قانون جهان است. اما هر چیز جدید که شما به دنبال آن می گردید، به اندازه یک یا دو کلیک از منابع عمومی که در جدول ۳-۱۵ لیست کرده ایم، فاصله دارد. در این لیست ۴

لینک آخر مربوط به منابع خیلی خوب آنلاینی هستند که تقریباً هر چیز مربوط به بیوانفورماتیک را می توانید در آن ها پیدا کنید.

Table 15-2 Ten Important Software Programs in Bioinformatics

Category	*Name*	*Address*	*Description*
Database Search	SRS	srs.ebi.ac.uk	Database search
	Entrez	www.ncbi.nih.gov/Entrez	Database search (Chapter 3)
	BLAST	www.ncbi.nlm.nih.gov/blast	Homology search (Chapter 7)
	DALI	www.ebi.ac.uk/dali	Structure database search (Chapter 11)
Multiple alignment	ClustalW	www.ebi.ac.uk/clustalw	Multiple sequence alignment (Chapter 9)
	MUSCLE	phylogenomics.berkeley.edu/muscle/	Multiple sequence alignment (Chapter 9)
	Tcoffee	www.tcoffee.org	Multiple sequence alignment (Chapter 9)
Prediction	GenScan	genes.mit.edu	Gene prediction (Chapter 5)
	PsiPred	bioinf.cs.ucl.ac.uk/psipred/	Protein structure prediction (Chapter 11)
	Mfold	www.bioinfo.rpi.edu/applications/mfold/	RNA structure prediction (Chapter 12)
Phylogenetics	Phylip	bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html	Tree reconstruction (Chapter 13)
	PhyML	atgc.lirmm.fr/phyml/	Tree reconstruction (Chapter 13)
Edition/ Visualization	Jalview	www.jalview.org	Alignment editor (Chapter 10)
	Logos	weblogo.berkeley.edu	A MSA Visualization Tool (Chapter 10)
	Trees	iubio.bio.indiana.edu/treeapp/treeprint-form.html	Tree Visualization (Chapter 13)
	Rasmol	www.umass.edu/microbio/rasmol/	Structure visualization (Chapter 11)

Table 15-3	Ten Bioinformatics Resource Locators	
Name	***Address***	***Description***
ExPASy	www.expasy.ch	Dedicated to proteins
ArrayExpress	www.ebi.ac.uk/microarray/	DNA chips
Swbic	www.swbic.org/	Miscellaneous links
Pasteur	bioweb.pasteur.fr/intro-uk.html	Miscellaneous links; many online tools
RNA World	www.imb-jena.de/RNA.html	RNA-related links
miRNAs	microrna.sanger.ac.uk/sequences/index.shtml	Extensive Resources on miRNA
Phylip	evolution.genetics.washington.edu/phylip/software.html	Everything on phylogeny
NCBI primers	www.ncbi.nlm.nih.gov/education	Very good primers on many subjects
Bielefeld	bibiserv.techfak.uni-bielefeld.de/intro/dist.html	Awesome online course
Bio-informer	www.ebi.ac.uk/Information/News/	The EBI online news
Coffee Corner	www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=coffeebrk	NCBI Online News

جاهایی که می توانید جریان های در حال وقوع را دنبال کنید

اگر همیشه دنبال گوشت تازه یا کیک های داغ می گردید، منابع ۴–۱۵ به درد شما می خورند. اگر می خواهید در بیوانفورماتیک صاحب نظر باشید، یک نگاهی به آن ها بیندازید.

Table 15-4	Ten Places to Go Farther	
Name	***Address***	***Description***
Nucleic Acid Research	nar.oxfordjournals.org/	Once a year, NAR publishes both a data-base issue and Web-server issue. These are available for free — and contain the state of the art in bioinformatics.
Bioinformatics	bioinformatics.oxfordjournals.org/	Bioinformatics contains articles describing the most recent methods in bioinformatics.
Conferences	www.iscb.org/events/event_board.php	An exhaustive list of major conferences in the field of bioinformatics, provided by the International Society For Computational Biology.