بسمه تعالی

بیوانفورماتیک

گروه بیوتکنولوژی دریا

کارشناسی ارشد

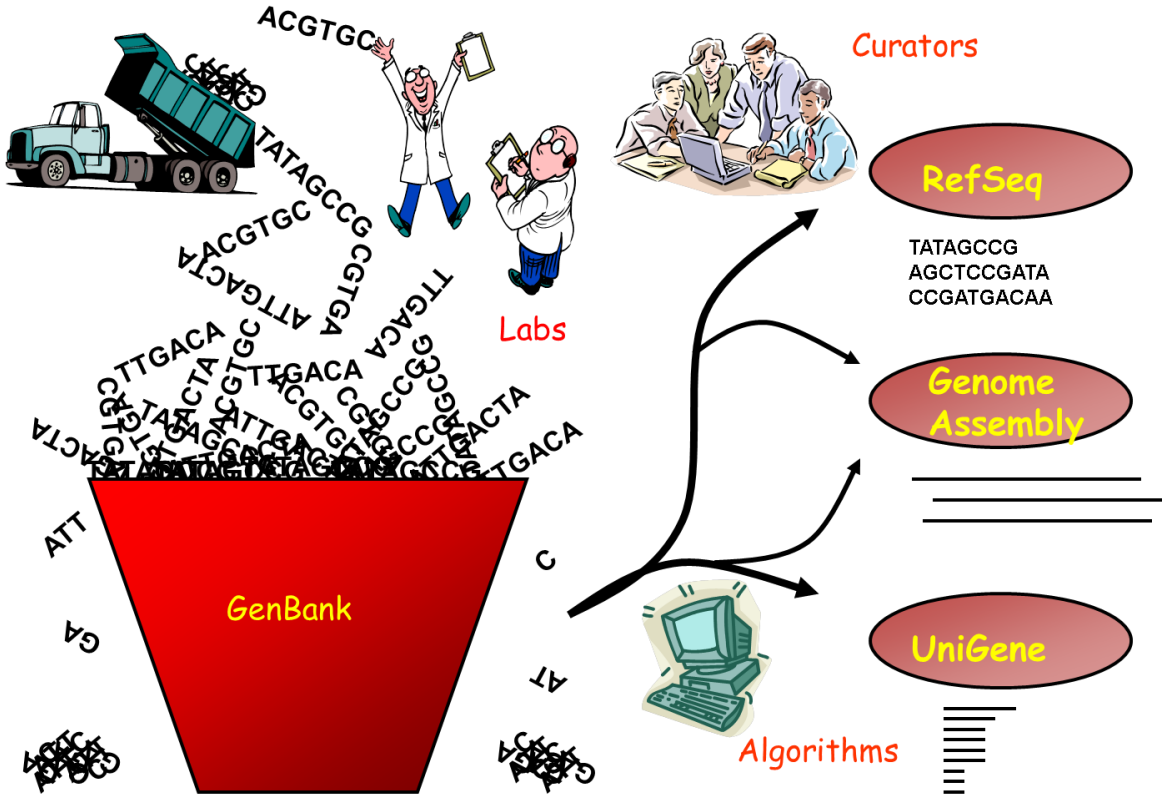
معرفی پایگاه‌های اطلاعاتی اولیه و ابزارهای جستجو در آنها

مقدمه

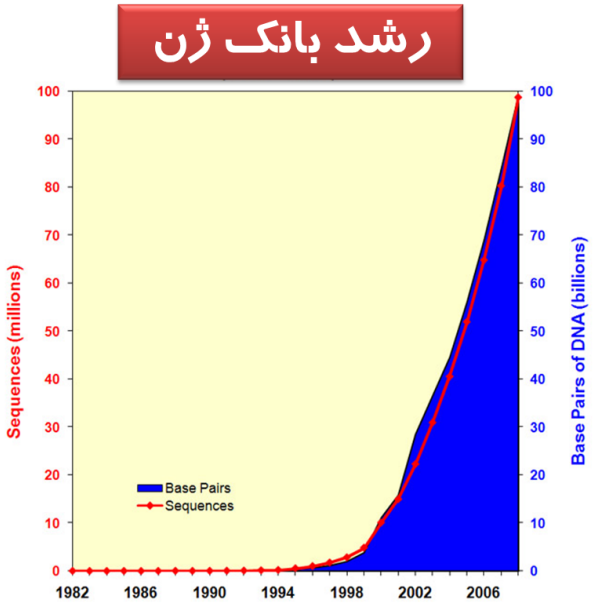
ما در دنیایی زندگی می کنیم که در آن حدود یک ونیم میلیون موجود زنده شناخته شده است و حدود ده هزار گونه آنها در خشکی‌ها زندگی می‌کنند. البته تعداد کل گونه‌های روی کره زمین ۱۰ تا ۱۰۰ برابر آن تخمین زده می‌شود. برای هر یک از این موجودات می‌توان داده‌های زیادی را جمع‌آوری نمود که شامل داده‌های ژنومی ، ژن‌ها، نحوه بیان آنها، ساخت پروتئین‌ها وتوالی آنها و مانند آن می باشد. از پردازش این داده ها، اطلاعات فراوانی به دست می‌آید که خود حجم زیادی دارند. مجموعه این اطلاعات آن هم برای برخی از موجودات موضوع میلیون‌ها رکورد در انواعی از بانک‌های اطلاعاتی است. چگونگی بازیابی این اطلاعات و بهره‌برداری از آنها در آموزش و پژوهش زیست‌فناوری بسیار ارزشمند خواهد بود.

داده‌پردازی زیستی از اشتراک فعالیت‌های مرتبط با تولید هر گونه اطلاعات زیستی (از جمله پروژه‌های ژنوم)، ایجاد پایگاه‌های اطلاعاتی و ابزارهای مربوط، تولید نرم‌افزارها و ساخت سخت‌افزارهای مورد نیاز حاصل می‌شود. بازیابی اطلاعات حاصل، تجزیه و تحلیل و بالاخره تفسیر آنها نتیجه این فعالیت‌هاست که تاثیر عمیقی در رشد و پیشرفت علوم زیستی و رشته‌های مرتبط داشته است.

در حال حاضر، با تلاش‌های هماهنگ جهانی بیش از ۱۰۰۰ پروژه ژنوم در حال اجرا بوده و بیش از ۱۳۰ پروژه ژنوم خاتمه یافته است. در نتیجه سرعت تولید اطلاعات ژنتیکی به بیش از ۱۰۰۰ جفت باز در ثانیه رسیده است. گفته می شود در زمان کنونی حجم اطلاعات در GenBank هر ۱۰ ماه دو برابر می‌شود. به طوری که در سال ۱۹۷۷ میلادی تنها توالی یک ژن (گلوبین خرگوش) را می‌دانستیم. اکنون بیش از ۸۰ میلیون رکورد در این بانک وجود دارد. شمایی از رشد اطلاعات در این بانک در شکل زیر آورده شده است.

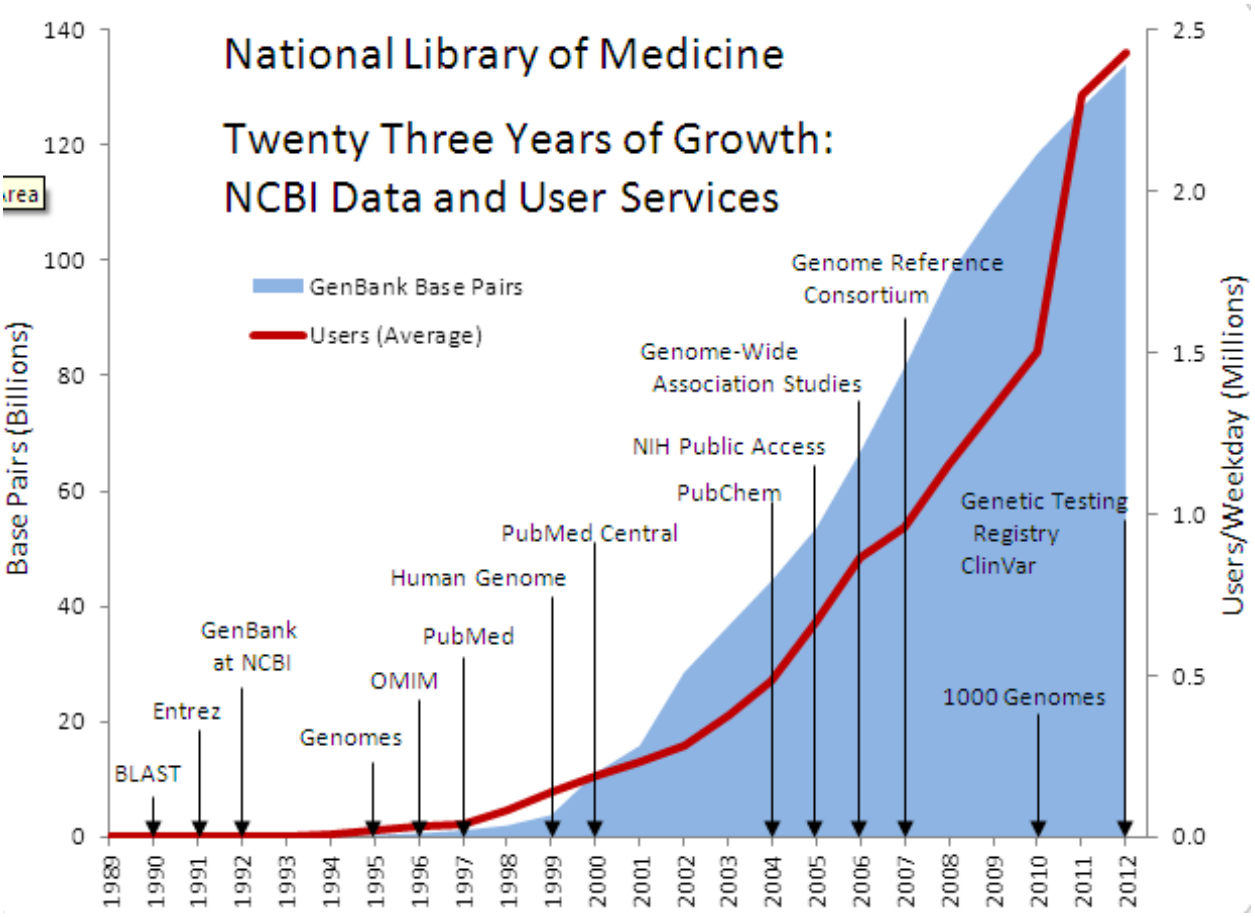


Year	Base Pairs	Sequences
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273
1995	384939485	555694
1996	651972984	1021211
1997	1160300687	1765847
1998	2008761784	2837897
1999	3841163011	4864570
2000	11101066288	10106023
2001	15849921438	14976310
2002	28507990166	22318883



شکل ۱. نمایشی از رشد اطلاعات در بیوانفورماتیک یا داده‌پردازی زیستی.

با کمی تاخیر زمانی، بهره‌برداری از این داده‌ها آغاز شده است. تشخیص مولکولی و زود هنگام بیماری‌ها در انسان، دام و حیوان، شناسایی ژن‌های مفید، طراحی موجوداتی با ویژگی‌های کیفی مورد نظر، اصلاح روش‌های زراعی و دامپروری مثال‌هایی از کاربردهای فعال این اطلاعات هستند. اما پیش‌بینی می‌شود در آینده نزدیک بهره‌برداری از این اطلاعات با مشابه سازی رایانه‌ای سامانه‌های زیستی وسعت بیشتری یابد. هم اکنون کشورهای پیشرفته با پایه‌گذاری رشته‌ای تحت عنوان زیست‌سامانه (System Biology) و تأسیس موسسات پژوهشی مرتبط درصددند تا سال ۲۰۱۵ به این هدف جامه عمل بپوشانند. انتظار می رود با دسترسی به سامانه‌های زیستی رایانه‌ای بسیاری از طراحی‌های فناوری و بررسی اثربخشی و مخاطرات آنها با هزینه بسیار کمتر و با سرعتی بسیار بیشتر (۴ تا ۵ برابر) انجام شود. به عبارت دیگر، در مقطعی از زمان قرار گرفته‌ایم که سرنوشت زیست فناوری در آن رقم می‌خورد. بدیهی است در این حرکت بسیار پرشتاب آشنایی با داده پردازی زیستی و پروژه های ژنوم مبنای ورود در این عرصه است



شکل ۲. نمایشی از رشد اطلاعات در بیوانفورماتیک.

پایگاه داده چیست؟

پایگاه داده‌ها یا بانک اطلاعاتی به مجموعه‌ای از داده‌ها با ساختار منظم و سامانمند گفته می‌شود. به عبارت ساده‌تر، پایگاه اطلاعاتی مجموعه‌ای از اطلاعات رکوردهای مرتبط می‌باشد و این اطلاعات طوری تنظیم شده‌اند که توانایی دستیابی به هر یک از قسمتهای اطلاعات مورد نظر امکان‌پذیر است (تعریف ارائه شده درسایت کتابخانه مک کونیل).

پایگاه داده‌ها در علوم زیستی

به دلیل افزایش سریع داده‌های زیستی اعم از مقالات، ثبت اختراعات، گزارش‌ها، توالی‌های نوکلئوتیدی یا اسید آمینه‌ای، ساختارهای سه بعدی، بررسی‌های ژل الکتروفورز و... ایجاد پایگاه های زیستی ضرورت یافته است. یکی از اولین پایگاه‌ها GenBank بوده است که در سال ۱۹۸۲م با تنها ۵ توالی شروع شد. این پایگاه هم اکنون میلیون‌ها رکورد در خود دارد.

دستجات مختلفی از پایگاه های اطلاعاتی زیستی وجود دارد. از جمله:

- پایگاه‌های اطلاعاتی اولیه شامل توالی های DNA و پروتئین،
- پایگاه‌های اطلاعاتی ثانویه شامل اطلاعات استخراج شده از پایگاه‌های اطلاعاتی اولیه،
- پایگاه‌های اطلاعات ژنوم‌ها،
- پایگاه‌های پلی مورفیسم/جهش،
- پایگاه‌های ساختمان‌های سه بعدی،
- پایگاه‌های ساختمان‌های سه بعدی پروتئین‌ها
- پایگاه‌های اطلاعاتی متابولیزم

پایگاه اطلاعات اولیه داده‌های زیستی

این پایگاه‌ها حاوی اطلاعات توالی‌های DNA و پروتئین است که به طور مستقیم از نتایج آزمایشگاهی حاصل شده‌اند (داده‌های تجربی یا Experimental) و یا از تجزیه و تحلیل رایانه‌ای داده‌های موجود به دست آمده‌اند (داده‌های پیش‌بینی شده یا Predicted). در این نوع پایگاه‌ها مولکول‌های فوق با علائم یک حرفی شاخص نوکلئوتیدها و اسید های آمینه نوشته می‌شوند (جداول زیر).

پایگاه‌های اطلاعات مربوط به کربوهیدرات ها و لیپیدها و اطلاعات ساختاری نیز گروهی دیگر از این پایگاه‌ها هستند که در حال توسعه هستند.

جدول کدهای نوکلئوتیدی

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

جدول کد ژنتیکی استاندارد

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

پایگاه‌های اطلاعات اولیه توالی DNA

این پایگاه‌ها در یک مسیر کاملاً سازماندهی شده مقادیر قابل توجهی از اطلاعات توالی هـای نوکلئوتیـدی کـه در آزمایشـگاه هـای مختلف تولید می شود را جمع آوری کرده و ذخیره می‌کنند. هر پایگاه اطلاعاتی قالب خاص خود را داشته و روش جستجو در آن نیز در عین سادگی از اختصاصیت برخوردار است.

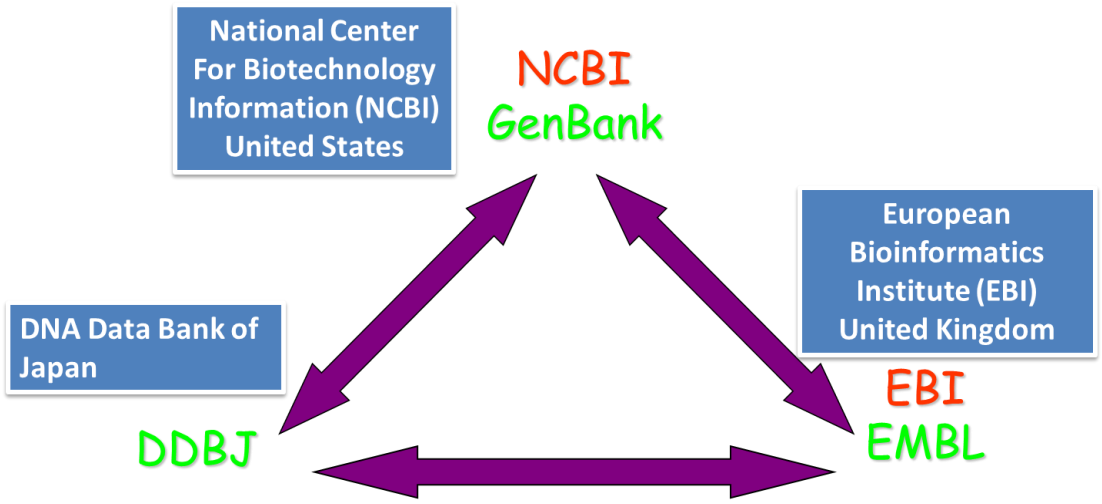
در جهان سه پایگاه اصلی برای حفظ، ذخیره و بازیابی اطلاعات نوکلئوتیدی (DNA و RNA) وجود دارد:

پایگاه GenBank در آمریکای شمالی که توسط مرکز ملی اطلاعات بیوتکنولوژی (NCBI) سرپرستی می‌شود.

پایگاه EMBL در اروپا که توسط آزمایشگاه زیست‌شناسی مولکولی اروپا (EBI) سرپرستی می‌شود.

پایگاه DDBJ در آسیا که توسط ژاپن ایجاد شده است.

در حال حاضر، این سه پایگاه در قالب یک معاهده همکاری به طور روزانه داده‌های جدید خود را مبادله می‌کنند.



شکل ۳. نمایشی از سه پایگاه نوکلئوتیدی عمده جهان و ارتباط بین آنها.

البته پایگاه‌های بسیار متنوع دیگری وجود دارند که مشتمل بر اطلاعات نوکلئوتیدی هستند و معمولاً در سه پایگاه فوق نیز وجـود دارند.

بانک‌های اطلاعاتی توالی حاوی تعداد زیادی فایل متنی (Text file) بسیار بلند هستند که اطلاعات مختلف مرتبط با توالی را در خود جای داده‌اند. حداقل اطلاعات موجود در یک فایل متنی عبارت است از:

- Header: اطلاعات ویژه یک توالی در آن قرار گرفته است؛
- Features: در این بخش Annotation های یک رکورد نمایش داده می‌شود؛
- Sequence: توالی را نشان می‌دهد.

داده‌های بانک‌های اطلاعاتی نوکلئوتیدی خود به چند گروه تقسیم می شوند:

- داده‌های mRNA یا cDNA موجود در GenBank، EMBL، DDBJ و...
- داده‌های DNA ژنومی: HTG، dbGSS و....
- داده‌ها dbEST: EST، uniGene و ...
- سایر داده‌ها: dbSTS، uniSTS، dbSNP، ...

لازم به ذکر است که توالی‌های RNA نیز به صورت DNA ذخیره می‌شوند. در هر دو حالت تنها یک رشته ذخیره می‌شود که رشته مثبت نامیده می‌شود.

توجه: رشته مثبت لزوماً حاوی توالی Sense نیست

بانک ژن (GenBank)

NCBI GenBank مشهورترین پایگاه داده‌های توالی نوکلئوتیدی و مستندات مربوط است که به عنوان بخشی از کتابخانه ملی پزشکی (National Library of Medicine) در سال۱۹۸۲ پایه‌گذاری شد. ابزارهای موجود برای دسترسی به داده‌های این پایگاه عبارتند از : BLAST (1990)، Entrez (1992)، GenBank (1992) و PubMed (1997). به دلیل تسلیم انواع داده‌های ژنومی، رشد اطلاعات در این بانک بسیار سریع بوده است (شکل ۱ و ۲). به طور میانگین، ماهانه ۳میلیون توالی و ۱۴۰۰ گونه جدید به این بانک اطلاعاتی افزوده می‌گردد به طوری که تقریبا هر ۱۰ ماه حجم اطلاعات آن دوبرابر می‌شود.
همان طور که در بالا گفته شد، داده‌های این پایگاه با بانک داده‌های DNA ژاپن (DDBJ) و آزمایشگاه بیولوژی مولکولی اروپا (EMBL) در حال تبادل بوده و هر سه پایگاه اطلاعات خود را روزانه رد و بدل می‌کنند.

توجه: GenBank حاوی داده های تکراری و اضافی (مانند توالی های وکتوری) است. بخشی از آن به دلیل تسهیل ورود اطلاعات بوده است اما دلیل عمده آن مجاز نمودن تسلیم توالی‌های مشابه تکراری از سوی آزمایشگاه‌های مختلف بوده است تا بتوان ذخیره‌ای از گوناگونی ژنتیکی (polymorphism) ایجاد نمود.

بخش های عمومی GenBank بر حسب نوع موجود که به بخش‌های تاکسونومیک مشهورند عبارتند از:

BCT (Bacterial and Archeal)
MAM (Mammalian)
Inv (Invertebrate)
PHG (Phage)
PLN (Plant and Fungi)
PRI (Primate)
ROD (Rodent)
SYN (synthetic=cloning vectors)
VRL (Viral)
VRT (other vertebrate)

GenBank همچنین به دلایل فنی (ماهیت حجیم و با کیفیت پایین داده‌ها) بخشی از اطلاعات خود را در قسمت‌های اختصاصی زیر ذخیره می‌کند:

PAT (patents)
EST (Expressed sequence tag)
STS (sequence tagged site)
GSS (Genome survey sequence)
HTG (High throughput genome)
CON (Contig)

اجزای یک رکورد GenBank

هر رکورد معمولی GenBank از سه قسمت تشکیل شده است:

Header: اطلاعات ویژه یک توالی در آن قرار گرفته است؛

Features: در این بخش Annotation های یک رکورد نمایش داده می شود؛

Sequence: توالی رکورد را نشان می‌دهد.

بخش عنوان یا Header

این بخش از رکورد حاوی اطلاعاتی نظیر نام لوکوس (که گاهی همان شماره دسترسی است)، نوع مولکول، زمان تسلیم اطلاعات، تعریف حاوی نام توالی، شماره دسترسی، شماره نسخه (شماره دسترسی به اضافهٔ یک عدد برای نشان دادن دفعات اصلاح آن رکورد)، منشاء توالی و تاکسونومی موجود منشاء، کلید واژه‌ها، آدرس مقالات مربوط می‌باشد. شماره دسترسی مهمترین جزء این قسمت است زیرا تنها این شماره یگانه بوده و بهترین گزینه برای مراجعه مجدد به این رکورد می‌باشد.

بخش ویژگی‌ها یا Features

در این بخش اطلاعاتی که در مورد ویژگی‌های توالی است می‌آید. طول توالی، محل توالی کدکننده پروتئین (CDS[1]) و اگزون‌ها (در صورت وجود)، گوناگونی در توالی (Variation)، توالی پروتئین رمز شده توسط آن توالی نوکلئوتیدی و مانند آنها در این بخش فهرست می‌شوند.

توالی یا Sequence

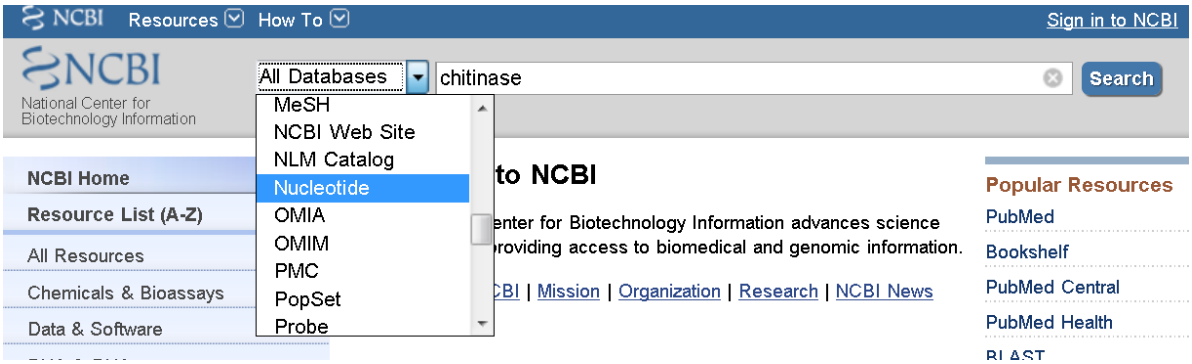
با ارائه آماری از انواع نوکلئوتید و پس از کلمه Origin توالی شروع می‌شود. اعداد در سمت چپ توالی موقعیت باز کنار آن در طول توالی را نشان می دهد. توالی ممکن است به کوچکی یک اولیگونوکلئوتید (بیش از ۲۰ جفت باز) در قسمت STS تا به بزرگی میلیون‌ها باز (توالی‌های کروموزم‌های به دست آمده از پروژه‌های ژنوم) باشد.

جستجوی متنی توالی نوکلئوتیدی

برای جستجوی متنی توالی‌ها مراحل زیر را انجام دهید:

وارد پایگاه NCBI شوید (http://www.ncbi.nlm.nih.gov) شوید و بخش All Database>Nucleotide را انتخاب کنید.

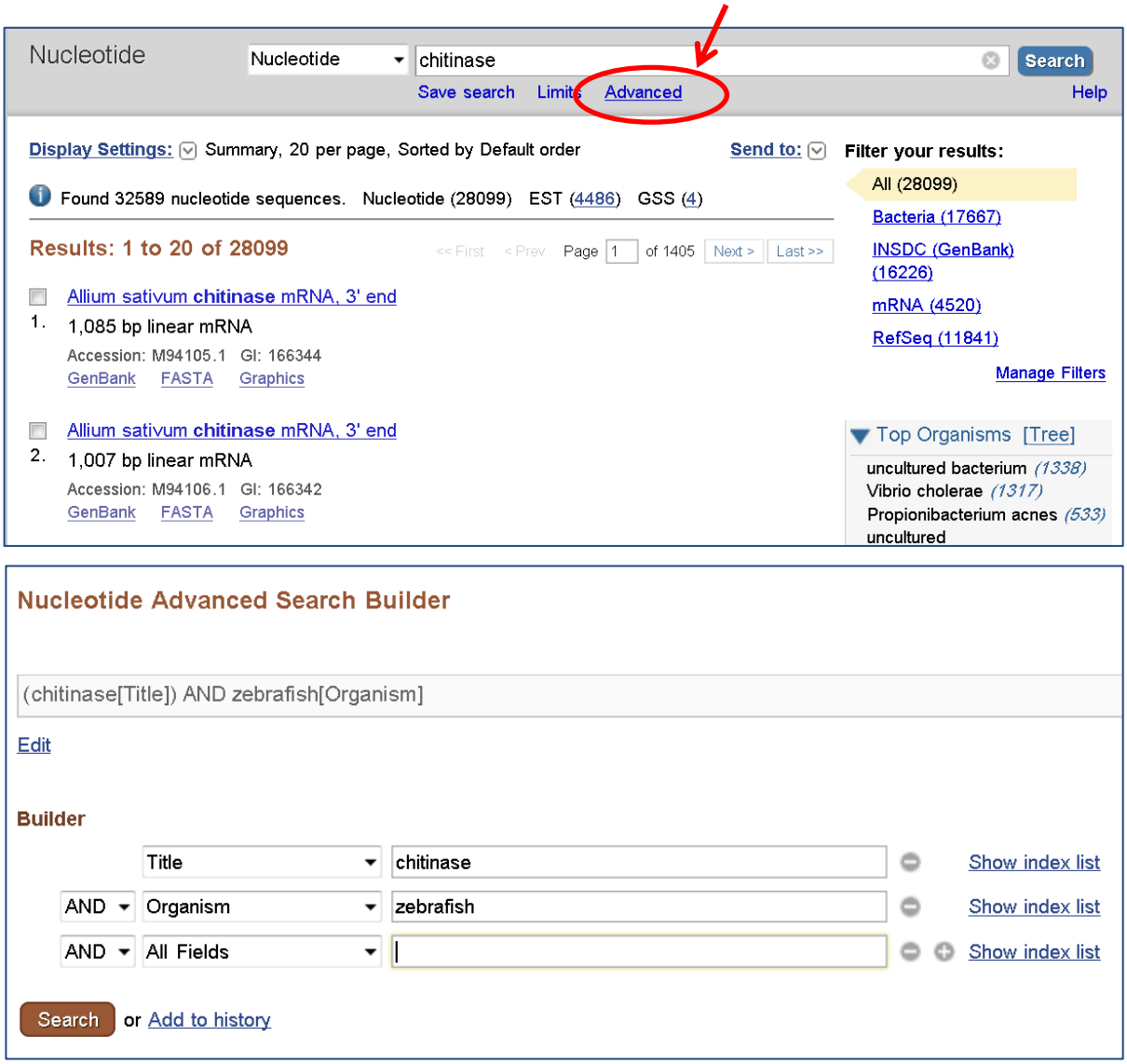
برای مثال واژه Chitinase را در محل جستجو بنویسید و روی دکمه Search کلیک کنید (شکل ۴) .



شکل ۴. جستجوی متنی توالی. برای یافتن توالی نوکلئوتیدی ژن chitinase این نام را در بانک Nucleotide جستجو می‌کنیم.

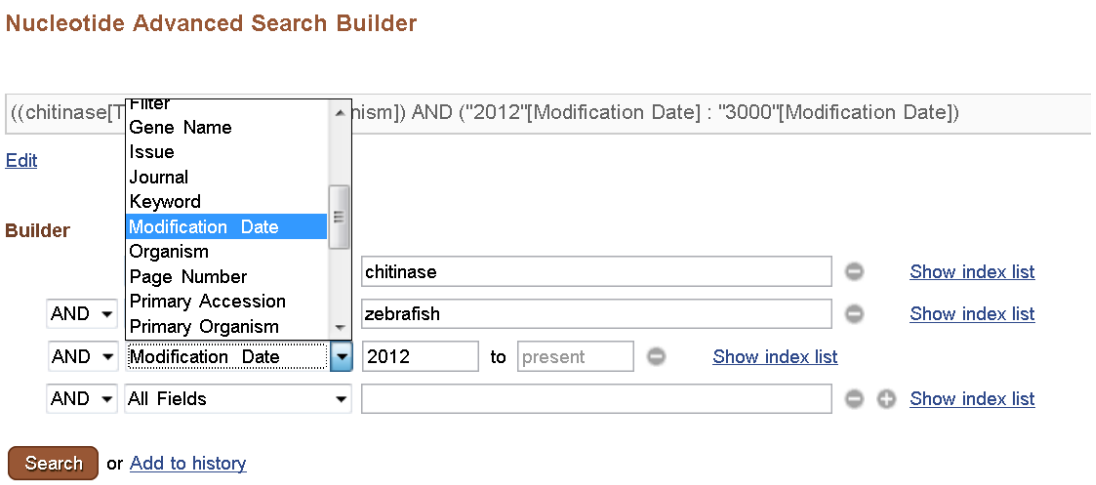
به دلیل زیاد بودن نتایج جستجو می‌توانید نتیجه را برای توالی مورد نظر خود محدود کنید برای این کار به قسمت Advanse Search وارد شوید. شما می‌توانید chitinase را در قسمت Title وارد کرده و اگر به دنبال توالی این ژن در یک نوع ماهی به نام Zebrafish (*Danio rerio*) می‌گردید، نام موجود مورد نظر را در قسمت Organism وارد کنید (شکل ۵).

[1] Coding sequence

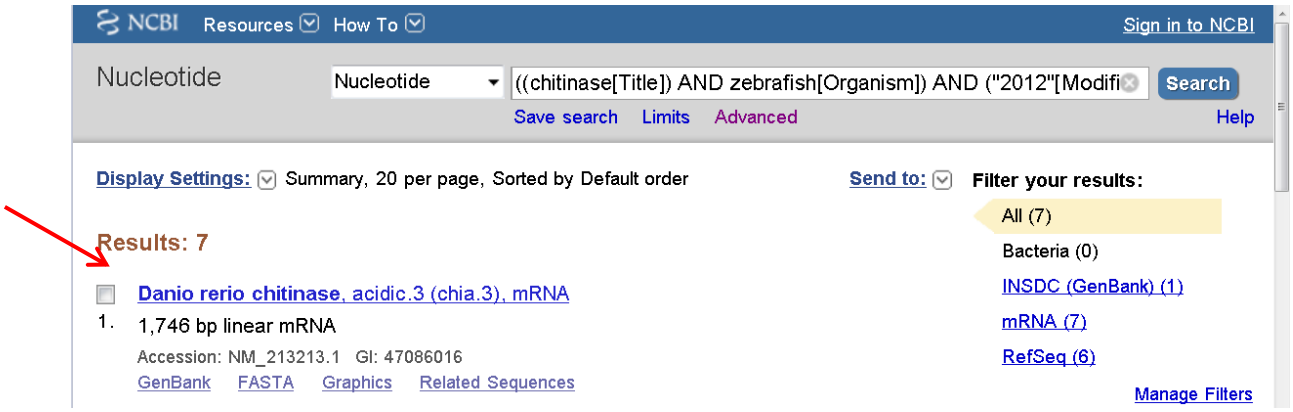


شکل ۵. جستجوی پیشرفته در بانک نوکلئوتید.

برای محدود کردن موارد دیگر می‌توانید از AND، OR یا NOT استفاده کنید. به عنوان مثال اگر بخواهید Chitinase را در موجود Zebrafish جستجو کنید به طوری که اطلاعات آن مربوط به بعد از سال ۲۰۱۲ باشد با استفاده از AND از پنجره مقابل گزینه Modification Date را انتخاب کرده و تایخ مورد نظر (در اینجا ۲۰۱۲) را تایپ کنید و عمل Search را انجام دهید (شکل ۶ و ۷).



شکل ۶. محدود کردن جستجو به موارد دیگر. در این شکل جستجو به توالی‌هایی که از سال ۲۰۱۲ تاکنون در مورد chitinase در Zebrafish به بانک نوکلئوتیدی NCBI وارد شده، محدود شده است.

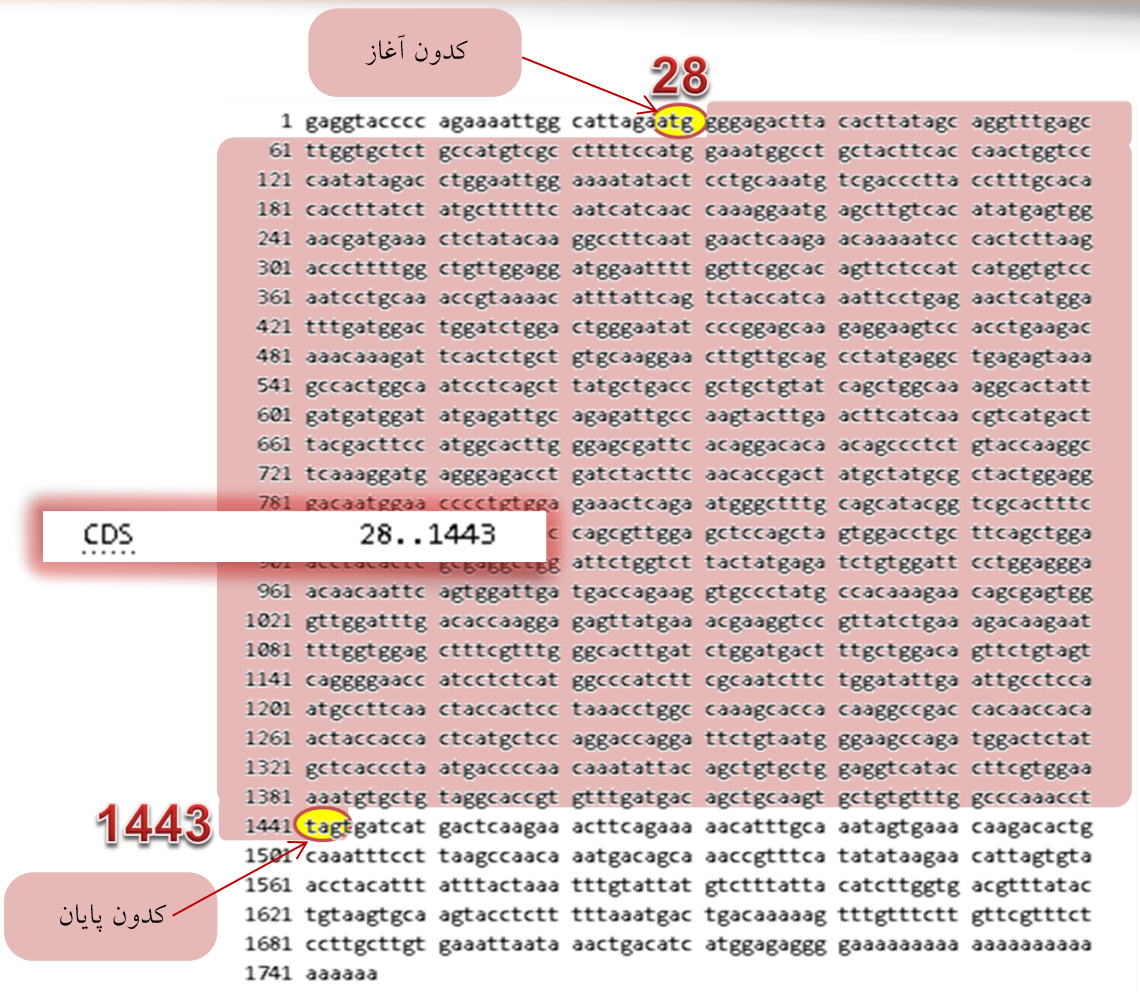


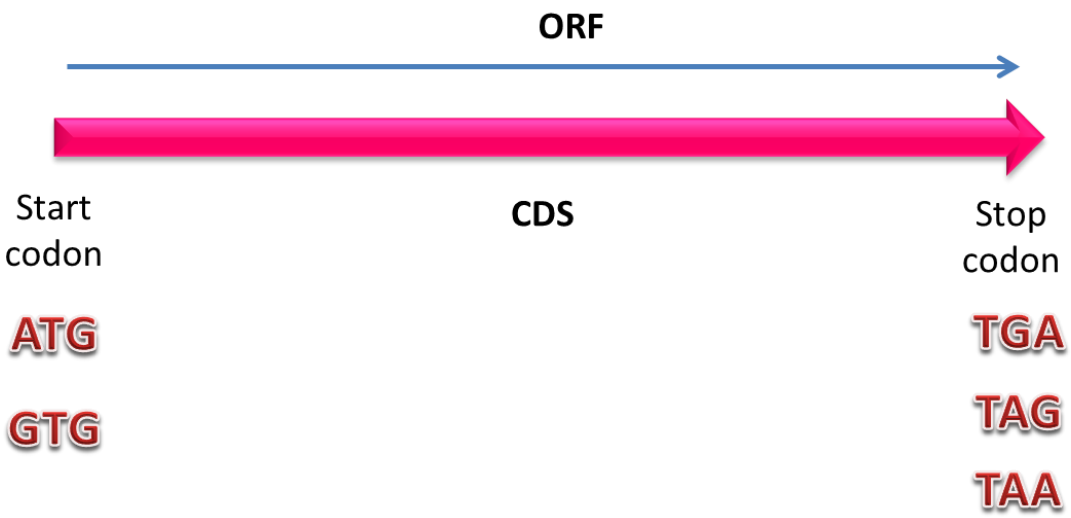
شکل ۷. نتیجه جستجوی محدود شده.

مشاهده می‌کنید که از ۲۸۰۹۹ مورد نتیجه جستجو در شکل ۵ به ۷ مورد نتیجه جستجو در شکل ۷ رسیدیم. اگر روی اولین نتیجه جستجو کلیک کنید صفحه مربوط به توالی نوکلئوتیدی مربوطه (شکل ۸) برای شما باز خواهد شد.

شکل ۸ صفحه نمایش یک توالی نوکلئوتیدی.

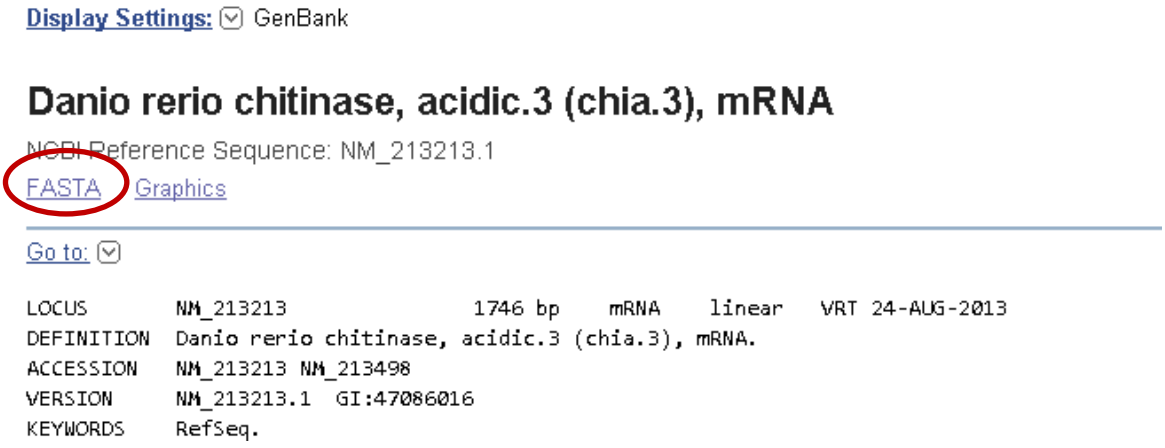
```
gene            1..1746
                /gene="chia.3"
                /gene_synonym="chioIb; wu:fd61a02; wu:fi13f01; zgc:55406;
                zgc:77912"
                /note="chitinase, acidic.3"
                /db_xref="GeneID:406819"
                /db_xref="ZFIN:ZDB-GENE-040426-2891"
CDS             28..1443
                /gene="chia.3"
                /gene_synonym="chioIb; wu:fd61a02; wu:fi13f01; zgc:55406;
```



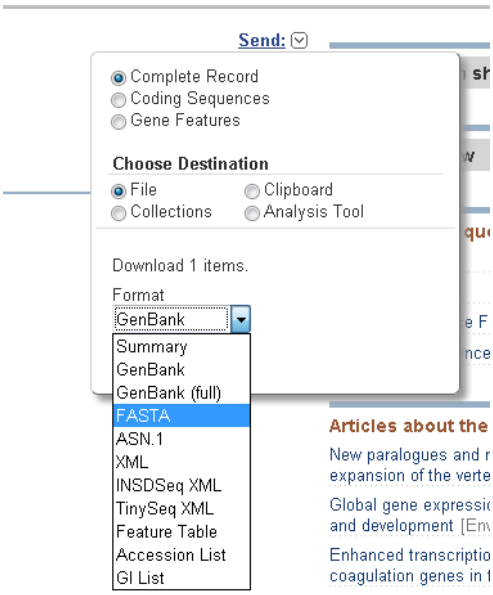


قالب‌های (Formats) نمایش توالی

پس از جستجو، می توانید رکورد یا توالی مورد نظر را به اشکال مختلف نمایش دهید. این موضوع هنگـامی اهمیـت مـی یابـد کـه می‌خواهید تجزیه و تحلیل بیشتر روی توالی به دست آمده با استفاده از نرم‌افزارهای دیگـر انجـام دهیـد. اکثـر نـرم‌افزارهـا قالـب FASTA را می پذیرند. در این قالب، سطر اول با علامت > آغاز شده و یک متن حاوی نام توالی به دنبال آن مـی آیـد. سـطرهـای بعدی فقط توالی هستند. از گوشه سمت چپ بالای صفحه نمایش توالی می‌توان به فرمت FASTA مربوط به توالی جسـتجو شـده دست پیدا کرد (شکل ۹). و مطابق شکل ۱۰ می‌توان یک توالی را با فرمت FASTA ذخیره کرد.



شکل ۹. دستیابی به فرمت FASTA برای توالی جستجو شده در NCBI.



شکل ۱۰. ذخیره توالی جستجو شده به فرمت FASTA (گوشه سمت راست بالای صفحه نمایش توالی در NCBI).

سایر انواع رکوردهایGenBank

۱- داده‌های EST: http://www.ncbi.nlm.nih.gov/dbEST/index

قسمت برچسب های بیانی توالی‌ها (Expressed sequence tags) حاوی توالی‌های کوتاهی است که معمولاً یک‌بار توالی‌یابی شده‌اند و مشتمل بر بخشی از توالی یک ژن هستند. این توالی‌ها همچنین از آزمایش‌های نمایش متفاوت (differential display) و RACE به دست می آیند.

۲- داده‌های GSS: http://www.ncbi.nlm.nih.gov/dbGSS/index

در این قسمت توالی های بررسی ژنومی (Genome Survey Sequences) ذخیره می شود. توالی های GSS کوتاه بوده و یک‌بار توالی‌یابی شده‌اند. این توالی‌ها تصادفی بوده و معمولاً از انتهاهای کلون‌های کاسمیدی و BAC به‌دست می‌آیند. توالی‌های مذکور در dbGSS ذخیره شده و از طریق قسمت GSS در GenBank نیز قابل دسترسی است. توالی و شماره دسترسی موجود در دو منبع، مشابه بوده ولی شکل رکورد متفاوتی دارند.

۳- داده‌های STS: http://www.ncbi.nlm.nih.gov/dbSTS/index

قسمت برچسب جایگاه توالی (Sequence tagged sites) حاوی توالی‌های کوتاه و یگانه در ژنوم است. از این توالی‌ها برای ایجاد نقشه‌های ژنتیکی استفاده می‌شود. این توالی‌ها از طریق dbSTS و بخش STS GenBank قابل دسترسی می‌باشند.

۴- داده‌های HTG: http://www.ncbi.nlm.nih.gov/HTG

این قسمت حاوی توالی‌های ژنومی است که توسط مراکز توالی‌یابی ژنومی در مقیاس وسیع به دست آمده است. این توالی‌ها از فازهای اتمام نیافته (صفر و ۱ و ۲) و اتمام یافته (فاز ۳) هستند. از داده‌های موجود در این قسمت می‌توان برای جستجوی BLAST بر علیه بانک داده‌های HTGs استفاده کرد که توالی آن در هر ماه به GenBank فرستاده می‌شود.

۵- داده‌های dbSNP: http://www.ncbi.nlm.nih/SNP

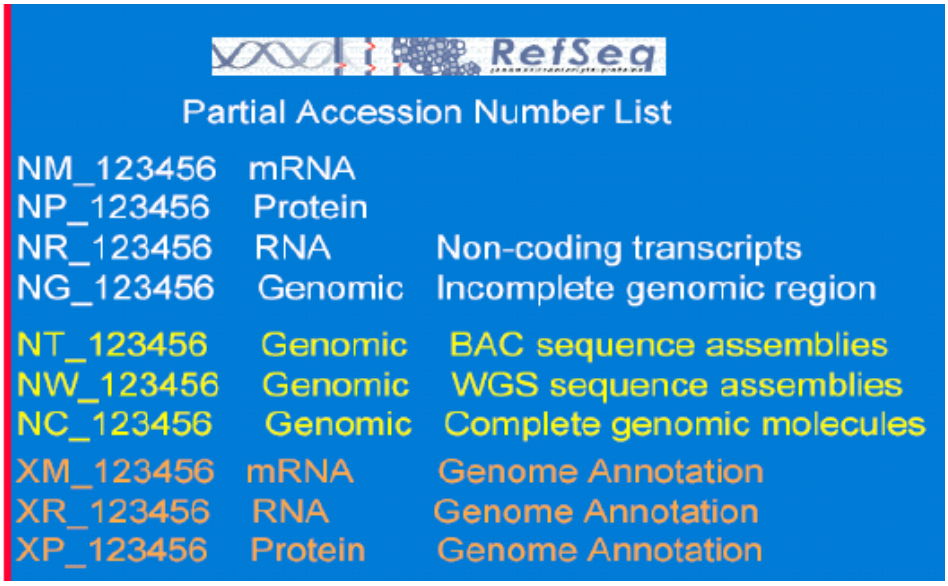
این پایگاه پلی مورفیسم های تک نوکلئوتیدی را ذخیره می کند. حذف ها و تداخل‌های هر توالی، تکرارهای پلی‌مورفیک و تنوع میکروساتلیتی در این پایگاه قرار می‌گیرد.

۶- داده‌های HTC: http://www.ncbi.nlm.nih/HTC

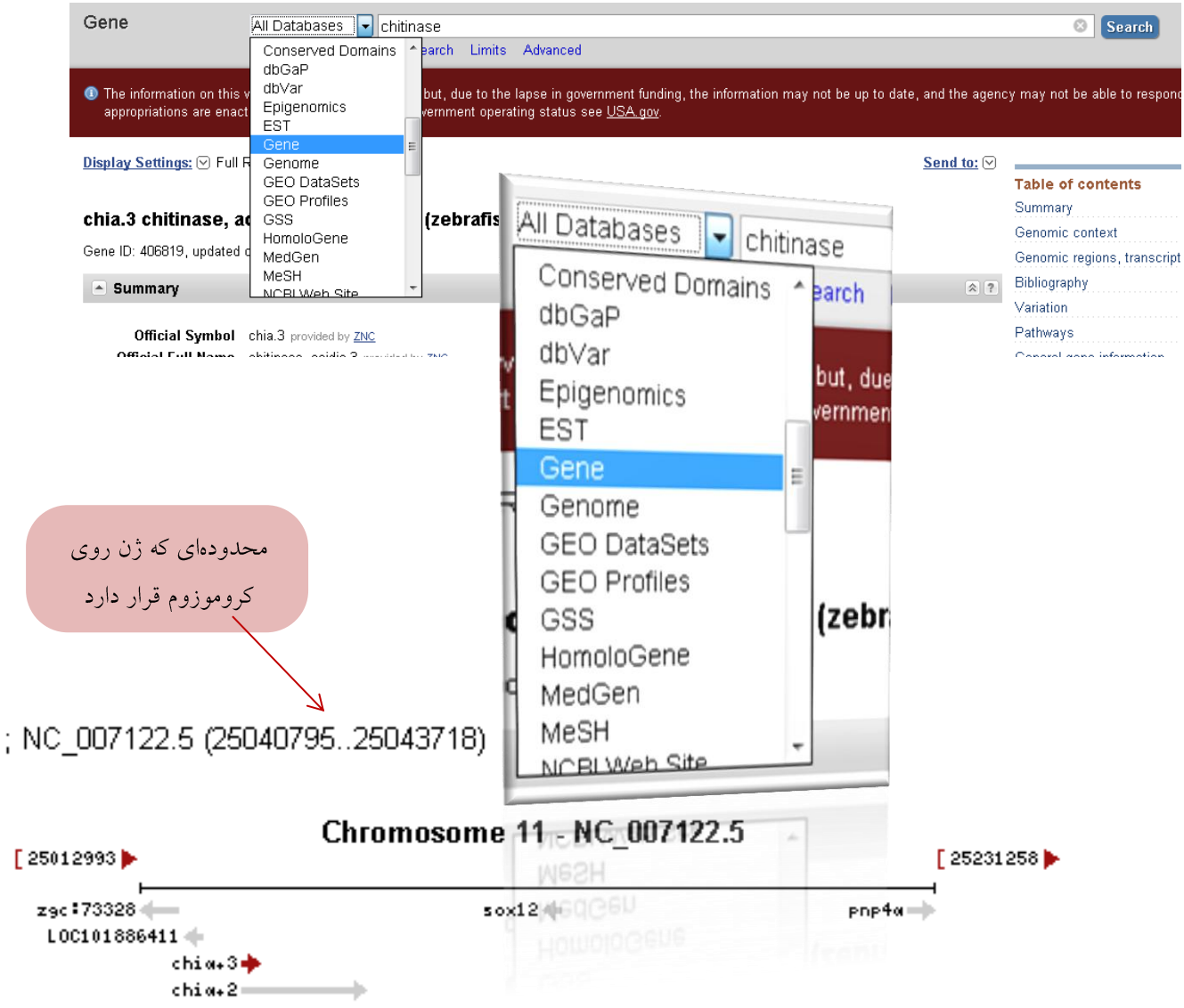
در ماه می سال ۲۰۰۰ سه پایگاه اطلاعاتی DDBJ، EMBL و GenBank برای ایجاد یک پایگاه جدید موافقت کردند. در این بخش توالی‌های cDNA قرار دارد که به صورت high throughput تولید شده‌اند و دارای بخش های 5'UTR، 3'UTR و یا بخشی از ناحیه رمز شونده (coding sequence) می‌باشند. بعد از اتمام توالی‌یابی HTCها، به بخش عمومی یا تاکسونومیک GenBank انتقال می‌یابند. توالی‌های HTC در ورود به بخش تاکسونومیک واژه کلیدی HTC را در ابتدای خود دارند. اما پس از ورود این واژه از ابتدای توالی حذف می شود.

پایگاه داده Refseq

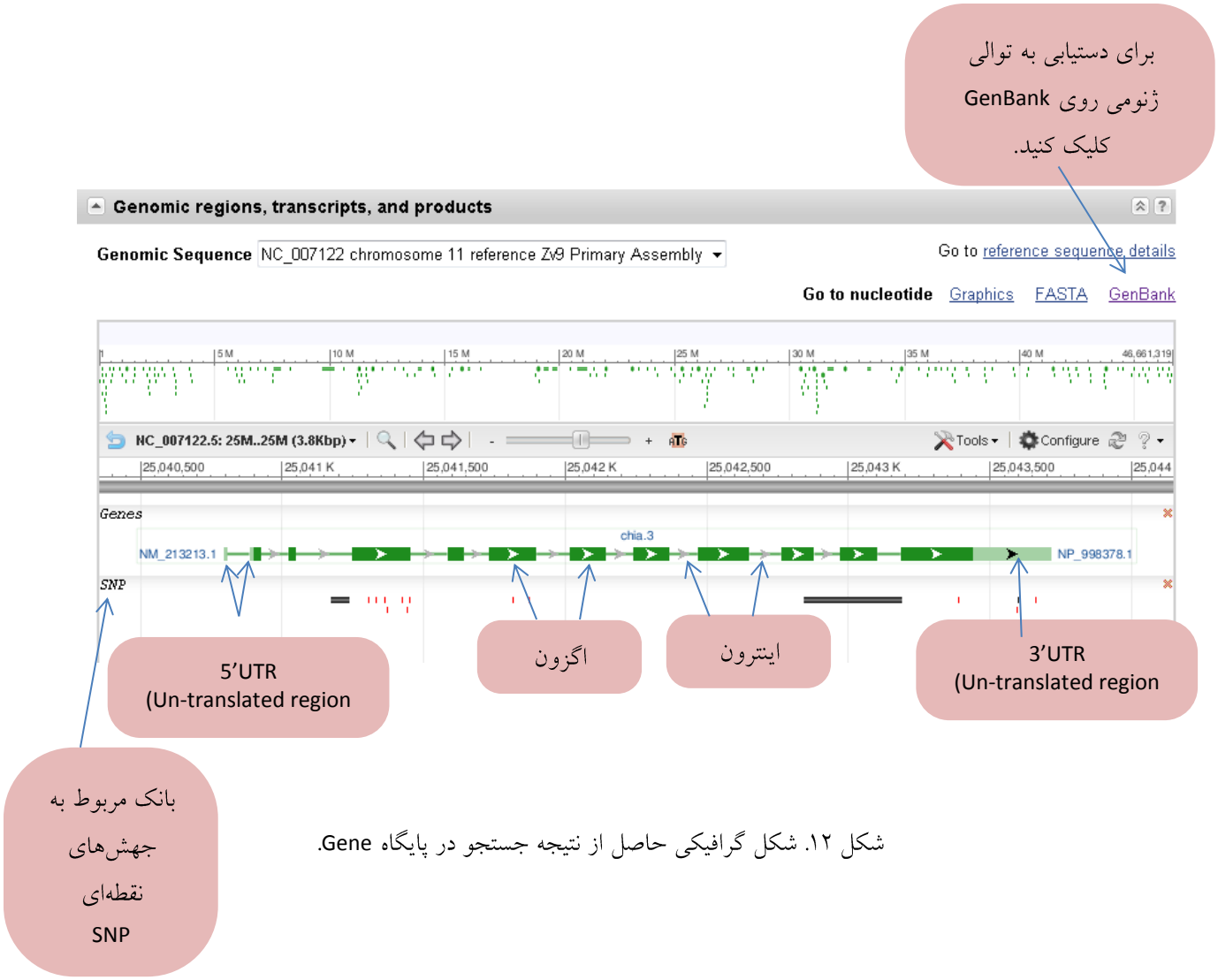
مجموعه توالی های مرجع استخراج شده از GenBank را گویند که تصحیح شده و غیر تکراری هستند. یعنی برخلاف GenBank، در پایگاهRefseq هر رکورد مربوط به یک ژن یا فرم پیرایش شده از یک ژن می‌باشد. در بانک داده‌های RefSeq شماره دسترسی هر مولکول با دستوری ویژه تعیین می‌شود که در شکل زیر مشخص شده است.



برای جستجوی مقالات می‌توان از بانک Pubmed، برای جستجوی توالی پروتئینی می‌توان از بانک پروتئین، برای داشتن اطلاعاتی در مورد موجود مورد نظر می‌توان از بانک Taxonomy، برای دستیابی به اطلاعات ژنوم‌ها از بانک ژنوم، برای دسترسی به داده‌های بیان ژن از قسمت Unigene، EST یا GEO می‌توان استفاده کرد. جستجو در قسمت Gene اطلاعات جالبی در مورد ژن از جمله شکل گرافیکی محدوده ژن، mRNA و پروتئین را نشان خواهد داد و محدوده اگزون‌ها و اینترون‌ها را نشان می‌دهد. شکل ۱۱ نمونه‌ای از این نوع جستجو را نشان می‌دهد.



شکل ۱۱. جستجو در پایگاه Gene.



شکل ۱۲. شکل گرافیکی حاصل از نتیجه جستجو در پایگاه Gene.

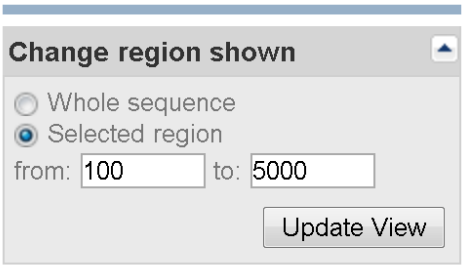
اگر بعد از جستجوی توالی در پایگاه Gene روی قسمت GenBank (شکل ۱۲) کلیک کنید به توالی ژنومی محدوده‌ای که مشخص شده دست پیدا خواهید کرد. در قسمت feature توالی‌های ژنومی نوکلئوتیدهایی که از اتصال آنها mRNA ایجاد می‌شود، مشخص شده است. به عنوان مثال در شکل ۱۳ قسمت mRNA مشخص می‌کند که از نوکلئوتید ۴۴۰ تا ۴۵۱ اگزون اول و از نوکلئوتید ۵۳۰ تا ۵۶۹ اگزون دوم وجود دارد و بنابراین نوکلئوتیدهایی ۴۵۲ تا ۵۲۹ اینترون بوده‌اند که در mRNA حذف می‌شوند. همانطور که در شکل ۱۳ مشاهده می‌کنید mRNA از نوکلئوتید ۴۴۰ شروع می‌شود، درحالی‌که شروع CDS نوکلئوتید ۵۴۵ است و این بدان معناست که از نوکلئوتید ۴۴۰ تا ۵۴۴ رونویسی می‌شوند اما ترجمه نمی‌شوند. به این قسمت 5'UTR یا ناحیهٔ بدون ترجمهٔ سمت 5' گفته می‌شود که می‌توان جایگاه اتصال ریبوزوم (RBS) را به همراه داشته باشد. بر همین اساس محدودهٔ 3'UTR بین نوکلئوتیدهای ۳۰۸۴ تا ۳۳۶۳ است.

```
gene            440..3363
                /gene="chia.3"
```

```
mRNA            join(440..451,530..569,667..693,894..1098,1229..1285,
                1378..1543,1663..1787,1888..2011,2113..2295,2408..2524,
                2615..2747,2830..3363)
                /gene="chia.3"
```

```
CDS             join(545..569,667..693,894..1098,1229..1285,1378..1543,
                1663..1787,1888..2011,2113..2295,2408..2524,2615..2747,
                2830..3083)
                /gene="chia.3"
```

شکل ۱۳. نحوه نمایش ویژگی (Feature) mRNA و CDS در یک ژن حاوی اگزون و اینترون (توالی ژنومی).



محدوده نمایش توالی‌های ژنومی را می‌توان بر حسب نیاز مطابق شکل بالا تغییر داد تا به توالی‌های اطراف توالی مورد جستجو دست پیدا کرد.

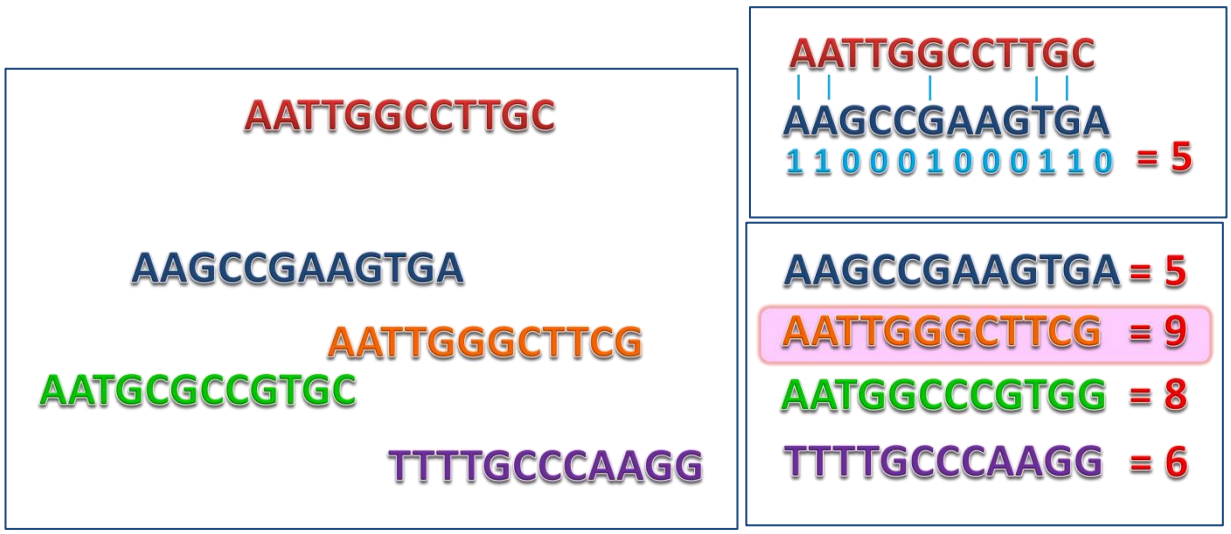
همردیفی دوگانه

مقایسه دو توالی

در دهه ۸۰ یک محقق هیچ برنامه رایانه‌ای برای این که بتواند بین تعدادی توالی، توالی‌ای را پیدا کند که بیشترین مشابهت را به توالی موردنظر خود داشته باشد در اختیار نداشت. بنابراین زبده‌ترین دانشمندان آن روزگار نیز مجبور بودند این کار را به صورت دستی انجام دهند. به طور مثال، اگر قرار بود از بین چهار توالی زیر مشابه‌ترین توالی را به توالی خود انتخاب کنند، باید تک تک توالی‌ها را با توالی الگو مقایسه کرده و میزان شباهت‌ها را در هر مقایسه به دست می آوردند. امروزه به این همردیفی دوگانه (Pairwise Alignment) می‌گویند.

ساده‌ترین راه برای مقایسه کردن دو توالی این است که هر بار دو توالی را در زیر هم قرار دهیم و یک به یک بازها را با هم مقایسه کنیم تا شبیه‌ترین توالی را پیدا کنیم (شکل ۱۴). ولی سوالی که پیش می‌آید این است که چگونه و با چه معیاری دو توالی مشابه‌تر را انتخاب کنیم؟ در این جاست که بحث امتیاز دهی (Scoring) مطرح می شود. به عنوان مثال، ساده ترین نوع امتیاز دهی این گونه می‌تواند باشد که اگر دو بازی که زیر هم قرار می گیرند یکسان باشند، امتیاز ۱ و اگر همسان نباشند، امتیاز صفر داده شود. با این روش می‌توان مشابه‌ترین توالی و درجه شباهت سایر توالی‌ها را با توالی الگو به دست آورد.

حال توالی‌های بالا را با توالی الگوی داده شده، با همین روش، همردیفی دو گانه می‌کنیم. در شکل ۱۴ به نظر می‌رسد توالی که امتیاز ۹ گرفته است، همسانی بیشتری با توالی ما دارد.



شکل ۱۴. همردیفی دوگانه و امتیازدهی به روش صفر و یک.

در هم‌ردیفی دوگانه مساله این است که دو توالی چقدر به هم شبیه هستند. زمانی که ما برای همسانی از امتیازدهی و عدد استفاده می‌کنیم، در واقع از روش‌های ریاضی برای حل مسأله زیستی استفاده می‌کنیم. از آنجا که در دنیای زیست‌شناسی، پارامترهای دخیل بسیار زیاد و در بسیاری از موارد ناشناخته هستند، بنابراین برای حل مسأله‌های زیستی با استفاده از الگوریتم‌های ریاضی و

رایانه‌ای، همواره با مشکل عدم تطبیق کامل مدل ریاضی با واقعیت زیستی روبرو هستیم. تفاوتِ راه حل‌ها و الگوریتم‌ها با هم در این است که جواب کدام یک به واقعیتِ زیستی که مشاهده می‌شود نزدیک‌تر است و آن را بهتر توجیه می‌کند.

اما همانطور که در شکل زیر دیده می‌شود، این دو توالی را به طریق دیگری هم می‌توان هم‌ردیف کرد. لذا سوالاتی هنوز باقی است مانند آن که آیا همردیفی دیگری ممکن است؟ کدام همردیفی بهتر است؟ همردیفی بهتر یعنی چه؟ کدام همردیفی گویای اتفاقات زیستی است؟ آیا در همردیفی‌ها روندهای تکاملی قابل ردیابی است؟ تا چه حد؟ و چگونه می‌توان همردیفی‌ها را در این راستا به کار گرفت؟

اینها سوالات عمیقی است که پایه‌های اساسی داده‌پردازی زیستی را تشکیل می‌دهند. لکن در این قسمت سعی بر آن است که با اصول هم‌ردیفی تا اندازه‌ای آشنا شوید تا بتوان از آن برای جستجوی توالی‌های مشابه و قضاوت در مورد میزان مشابهت و درک مفهوم خانواده‌های ژنی و پروتئینی استفاده نمود.

روش Dot Plot

به عنوان راهی برای شناسایی تمامی همردیفی‌های ممکن محققین، روشی گرافیکی به نام دات پلات (dot plot) به کار بردند. در این روش دو توالی به صورت عمود بر هم روی محور x ها و y ها در یک صفحه قرار داده می‌شوند و در هر نقطه‌ای که شبیه هم باشند عدد یک قرار داده می‌شود. اگر دو توالی کاملاً شبیه باشند در نهایت، از اتصال نقاط، یک خط اوریب بدون شکستگی را می‌توان از انتهای بالای سمت چپ صفحه به انتهای پایین سمت راست صفحه رسم کرد. هم‌ردیفی در واقع مشخص کردن رابطهٔ بین نوکلئوتیدهای یک توالی با توالی دیگر است. اگر دو توالی در مثال فوق را به صورت دات پلات در آوریم جدول زیر به دست خواهد آمد. اگر دور بیش از دو عدد ۱ به دنبال هم خط بکشیم، منظره زیر ظاهر خواهد شد. حال بر اساس این خطوط اریب می‌توان کلیه همردیفی‌های دو گانه را استخراج کرده و به صورت خطی نوشت. اگر در قسمتی از یکی از توالی‌هایی که تحت آنالیز دات‌پلات قرار گرفته است وارونگی قسمتی از توالی اتفاق افتاده باشد، آنگاه خط اریبی در قسمتی از توالی خواهیم داشت که در قطر مقابل این قطر اصلی قرار می‌گیرد.

روش Dot Plot

	A	A	T	T	G	G	C	C	T	T	G	C
A	1	1	0	0	0	0	0	0	0	0	0	0
A	1	1	0	0	0	0	0	0	0	0	0	0
T	0	0	1	1	0	0	0	0	1	1	0	0
G	0	0	0	0	1	1	0	0	0	0	1	0
G	0	0	0	0	1	1	0	0	0	0	1	0
C	0	0	0	0	0	0	1	1	0	0	0	1
C	0	0	0	0	0	0	1	1	0	0	0	1
C	0	0	0	0	0	0	1	1	0	0	0	1
G	0	0	0	0	1	1	0	0	0	0	1	0
T	0	0	1	1	0	0	0	0	1	1	0	0
G	0	0	0	0	1	1	0	0	0	0	1	0
C	0	0	0	0	0	0	1	1	0	0	0	1

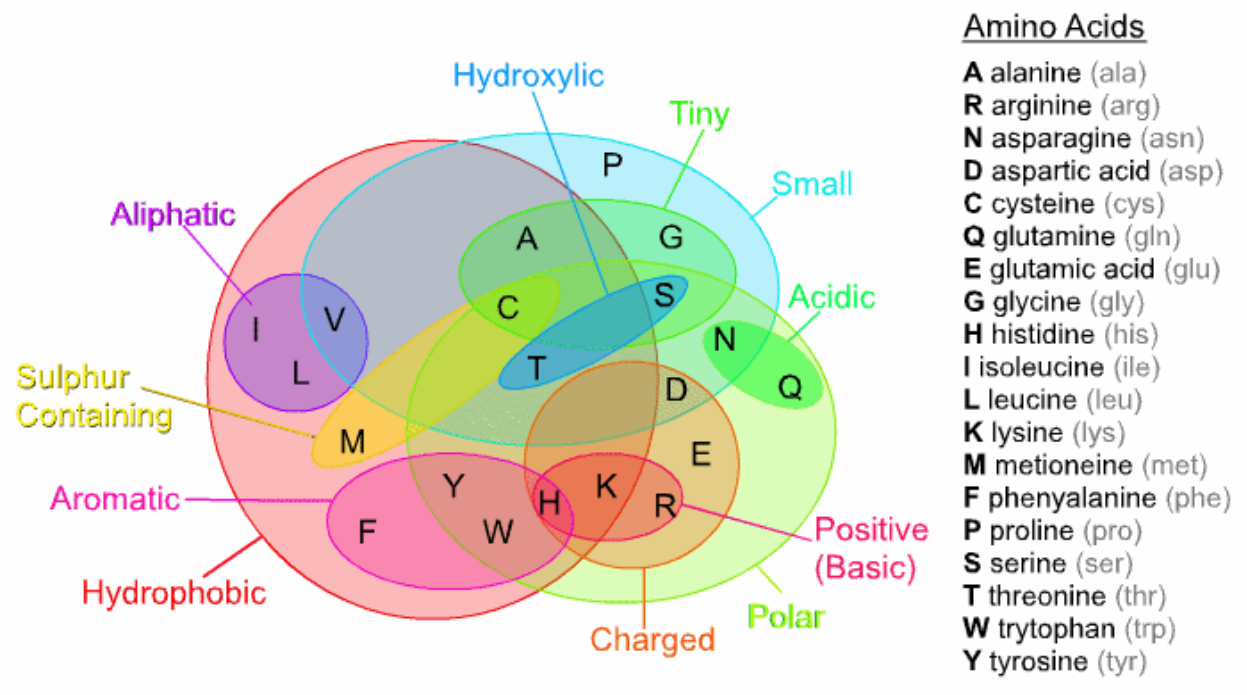
مقایسه همردیفی

در مثال‌های فوق روش امتیاز دهی صفر و یک را می‌توان یک نوع الگوریتم به حساب آورد که بر اساس آن هم‌ردیفی با بالاترین امتیاز را به عنوان بهترین هم‌ردیفی انتخاب کردیم. ولی در عمل می‌توان هم‌ردیفی‌هایی را مثال زد که با وجود امتیاز مساوی یا حتی بالاتر صحیح نبوده و با دانسته‌های قبلی تطبیق نمی‌کنند. بنابراین تلاش زیادی برای طراحی و به‌کارگیری الگوریتم‌هایی دقیق‌تر صورت می‌گیرد که تا هر چه بیشتر دربرگیرنده واقعیات زیستی و اصول حاکم بر حیات باشد.

به طور مثال، در همردیفی توالی‌های نوکلئوتیدی می‌توان بین امتیازات جایگزینی‌های از نوع جانشینی (Substitution) و انتقال (Transition) تفاوت قائل شد. زیرا با توجه به ساختمان دو رشته‌ایDNA احتمال جایگزینی بازهای پورینی با هم و بازهای پیریمیدینی با هم بیشتر است. در حالی که الگوریتم قبلی تفاوتی را بین این دو حالت قائل نبود. بنابراین جواب‌هایی هم که با الگوریتم قبلی به دست آمد کمتر به واقعیت نزدیک است.

این مشکل در توالی‌های پروتئینی به طور جدی‌تری مطرح است. در این توالی‌ها، همهٔ جایگزینی‌ها اشکال ساختاری و عملکردی زیادی ایجاد نمی‌کنند. به عبارت دیگر، برخی اسید آمینه‌ها خواص فیزیکوشیمیایی مشابهی دارند (شکل بعد) و می‌توانند با حداقل تغییر خواص جایگزین یکدیگر شوند.

نکتهٔ دیگر این است که در هم‌ردیفی‌های بالا فرض، برابر بودنِ طولِ توالی‌هاست. در حالی که در تکاملِ تـوالی‌هـا هـم پدیـده ی اضافه شدن را داریم و هم پدیده حذف اتفاق می‌افتد. بنابراین ممکن است در بسیاری از موارد بخواهیم دو توالی را با هم مقایسـه کنیم که دارای طول یکسانی نباشند.



جمع‌بندی این مقدمات نشان می‌دهد می‌توان با کمّی‌نمودن (امتیازدهی) نتایجِ هم‌ردیفی، آنهـا را بـاهم مقایسـه نمـود. البتـه بـرای کمّی‌نمودن همردیفی‌ها حداقل دو نوع امتیاز دهی را بایستی منظور کرد.

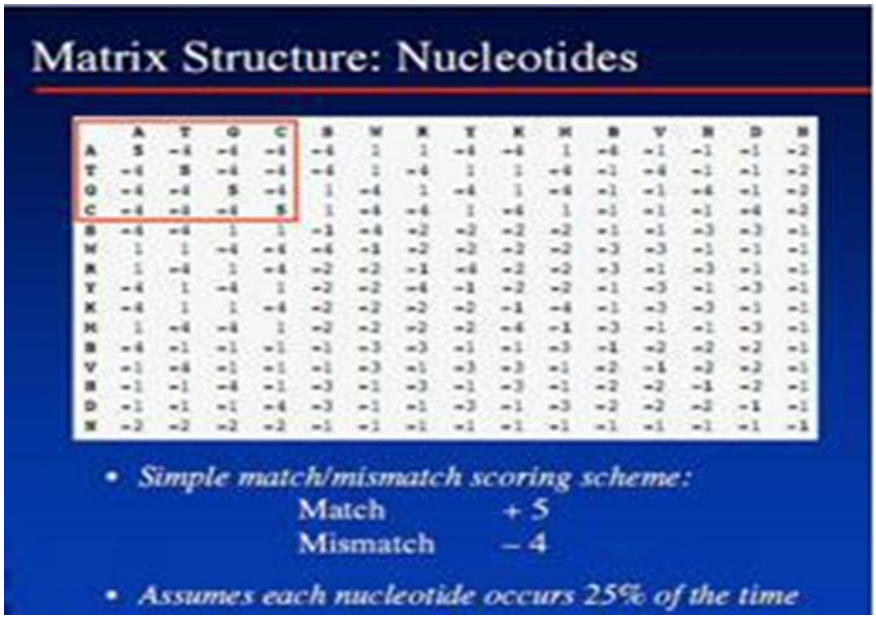
- **امتیاز دهی جایگزینی‌ها**
- **امتیاز دهی حذف و اضافه شدن توالی‌ها**

به این ترتیب، امتیاز هر هم‌ردیفی جمع جبری کلیه امتیازات جایگزینی‌ها و حذف یا اضافه‌ها خواهد بود.

امتیازدهی جایگزینی‌ها

با درک این که روش صفر ویک کفایت نمی‌کند و جایگزینی نوکلئوتیدها یا اسید آمینه‌ها با یکدیگر امتیاز منفی یـا مثبـت مسـاوی ندارند، متخصصینِ امر در پی تهیه جداول امتیازدهی جایگزینی‌ها (Substitution Scoring Matrices) بوده‌اند. به طوری که تا حـد امکان واقعیت‌های زیستی را منعکس نماید.

برای توالی‌های نوکلئوتیدی کار چندان دشوار نیست زیرا هر گونه جایگزینی منجر به جهش می‌شود که اثر آن در رمزدهی پروتئین‌ها ممکن است مشاهده شود. یعنی در این مولکول‌ها بحث ساختار و عمل چندان مطرح نیست. البته با توجه به ساختمان دو رشته‌ای DNA متخصصینِ تکامل زیستی بین جانشینی نوکلئوتید پورین و پیریمیدین و انتقال از پورین به پیریمیدین یا بالعکس تفاوت قائلند. جدول زیر نمونه‌ای از جداول امتیازدهی برای همردیفی دو توالی نوکلئوتیدی را نشان می‌دهد. در این جدول به طور ساده‌ای کلیه همسانی‌ها امتیاز ۵+ و برای غیر جفت شدگی امتیاز ۴- در نظر گرفته شده است. سایر حروف در صورت وجود انتخاب برای دو نوکلئوتید یا بیشتر در هر موقعیت از توالی کاربرد دارند .



در تهیه جداول امتیازات جایگزینی توالی‌های پروتئینی خواص فیزیکوشیمیایی اسیدهای آمینه و تاثیر جایگزینی آنها در ساختار و عمل پروتئین‌ها مطرح است. در گذشته، پژوهشگران به گروه‌بندی اسیدهای آمینه بر اساس خواص آنها (شکل صفحه قبل) مراجعه کرده و میزان مشابهت را به صورت توصیفی (و نه عددی) بیان می‌کردند. در دو دهه اخیر روش‌های تهیه جداول امتیازدهی جایگزینی، مبتنی بر داده‌های موجود در طبیعت بوده است. با فرض بر این که اگر دو اسید آمینه دارای خواص فیزیکوشیمیایی مشابهی هستند بایستی در طول تکامل جایگزینی آنها تحمل شده باشد، پژوهشگران نسبت به جمع‌آوری توالی‌ها، همردیفی آنها با هم و محاسبه فراوانی جایگزینی‌ها در بین پروتئین‌های هم‌خانواده اقدام نمودند.

در اولین تلاش، همردیفی ۱۵۷۲ توالی پروتئینی در ۷۱ درخت از ۳۴ خانواده پروتیئنی انجام و گروه‌بندی شد. سپس فراوانی جایگزینی یک اسید آمینه با اسید آمینه دیگر در فرمول زیر به‌کار گرفته شد :

PAM_n= nlog Probability of one substitution/Probability of occurring by chance*100

در این فرمول یک واحد PAM (Point Accepted Mutation) معادل یک تغییر در یک توالی صدتایی از اسیدآمینه‌هاست. داده‌های حاصل در جدول PAM ثبت می‌شود. از آنجا که در طول تکامل ممکن است اسیدآمینه در یک موقعیت چندین بار جایگزین شود، جدول حاصل را چندین بار در خود ضرب می‌کنند. به طور مثال، برای تهیه جدول PAM100 ان را ۱۰۰ بار در خودش ضرب می‌کنند (جدول زیر).

PAM
point accepted mutation

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0	-8
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1	-8
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	-8
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1	-8
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3	-8
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	-8
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	-8
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	-8
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1	-8
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-8
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	-8
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	-8
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	-8
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	-8
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1	-8
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0	-8
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	-8
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	-8
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2	-8
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	-8
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1	-8
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1	-8
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

بعدها جداول دیگری تدوین شدند که از یک نوع اصول پیروی می‌کردند. با این تفاوت که فراوانی جایگزینی‌ها تنها در مناطق حفاظت شده (Conserved Blocks) برای ساختن جدول، محاسبه می‌شدند. در آن هنگام، ۲۰۰۰ بلوک از ۵۰۰ خانواده پروتئینی در نظر گرفته شد. به طور مشابهی، فرمول زیر بکار رفت :

$BLOSUM_{\%}$=log Probability of substitution in block/Probability of occurring by chance

این جداول را BLOSUM نامیدند که از اصطلاحBlock Substitution Matrix برگرفته شده است. شماره جدول به نوع بلوک مورد استفاده برای محاسبه فراوانی و احتمال وقوع جایگزینی بستگی دارد. مثلا BLOSUM62 یعنی این جدول برمبنای فراوانی جایگزینی‌ها در بلوک‌های حاوی توالی‌هایی با همسانی ۶۲ درصد یا بیشتر تشکیل شده است (جدول زیر).

BLOSUM
(BLOcks SUbstitution Matrix)

BLOSUM62

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

چند اصطلاح:

همسانی یا Identity::

وقتی به کار می‌رود که منظور هم‌ردیفی یک نوکلئوتید یا اسیدآمینه با **همانند** آن منظور باشد.

مشابهت:

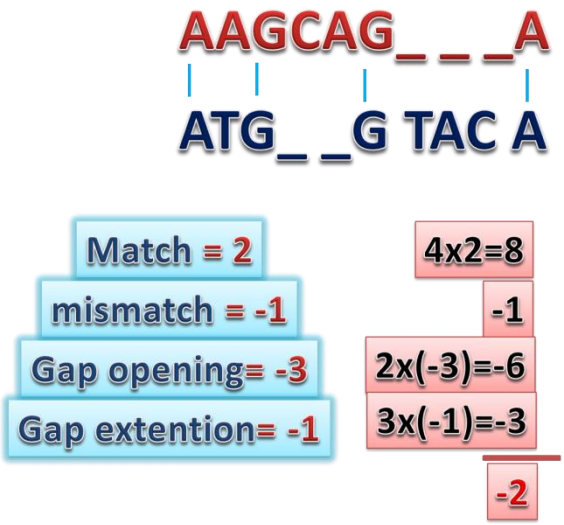
وقتی به کار می‌رود که منظور هم‌ردیفی یک نوکلئوتید یا اسیدآمینه با **معادل** آن منظور باشد.

همولوژی:

وقتی به کار می‌رود که منظور هم‌ردیفی یک نوکلئوتید یا اسیدآمینه با معادل آن منظور باشد و رابطهٔ نیاکانی بین توالی‌ها در نظر باشد.

امتیازدهی حذف و اضافه نمودن توالی ها

ممکن است جهش به صورت اضافه شدن یا حذفی توالی باشد. بروز این جهش ها در توالی‌ها باعث تفاوت در طول توالی‌ها می‌شود. بنابراین، زمانی که می خواهیم دو توالی را در حالت بهینه هم‌ردیف کنیم، نیازمند استفاده از فواصل هستیم. این فواصل بایستی به طریقی در محاسبه امتیاز یک همردیفی لحاظ شوند.



یک مثال از سیستم امتیازدهی فرضی.

این که فواصل را در همردیفی چگونه محاسبه کنیم یکی از مبهم‌ترین مساله‌ها در هم‌ردیفی توالی‌ها است. به طور معمول، جریمه‌هایی را که برای فواصل در نظر می‌گیرند به صورت محلی اعمال می‌شود. یعنی جریمه استفاده از هر فاصله مستقل از فواصل دیگری است که ممکن است در جاهای دیگری از هم‌ردیفی اتفاق بیافتد. در همهٔ برنامه‌ها، برای فواصل دو نوع امتیاز منفی در نظر گرفته می‌شود.

۱. **جریمه باز کردن توالی:(GOP[1])**

در طبیعت، هر اضافه نمودن توالی مستلزم صرف انرژی بوده و مورد انتخاب طبیعی قرار خواهد گرفت. بنابراین در الگوریتم‌ها برای هم‌ردیفی بهینه توالی‌ها امتیاز منفی نسبتاً بزرگی (مثلا ۱۱–) برای ایجاد فاصله در نظر گرفته می‌شود (در مثال بالا این امتیاز منفی ۳ در نظر گرفته شده بود).

[1]Gap opening penalty

۲. **جریمه بسط یک فاصله (GEP[1])**

از این جریمه برای ورود نوکلئوتید یا اسیدآمینه در محلی که قبلاً فاصله ایجاد شده است، استفاده می‌شود. میزان این جریمه از GOP کمتر فرض می شود ولی در تعداد آنها ضرب می‌شود (مثلا ۱- ضربدر تعداد). زیرا از دید زیستی جایی از توالی که شکافته شده است، استعداد ورود یک یا چند نوکلئوتید را دارد.

انواع هم‌ردیفی

همردیفی‌ها به دو شیوه قابل تقسیم‌بندی هستند:

الف- از نظر تعداد توالی

۱. **هم‌ردیفی دوگانه (Pairwise Alignment)**

هم‌ردیفی تنها دو توالی با یکدیگر در طول کامل آنها یا یک ناحیه خاص.

۲. **همردیفی چندگانه (Multiple Alignment)**

هم‌ردیفی سه یا چند توالی که از هم‌ردیفی‌های دو گانهٔ هر جفت آنها نتیجه می‌شود.

ب- از نظر طول توالی

۳. **هم‌ردیفی محلی (Local Alignment)**

یافتن و هم‌ردیفی بهترین محل‌های جور شدن دو توالی.

۴. **هم‌ردیفی کامل (Global Alignment)**

یافتن و هم‌ردیفی جورشدگی بین طول کامل دو یا چند توالی.

به طور خلاصه، اصول مطرح شده در صفحات قبل در تمامی انواع فوق استفاده می‌شوند. مثلاً هم‌ردیفی‌های چندگانه از جمع اطلاعات مربوط به کلیه هم‌ردیفی‌های دوگانه ممکن بین جفت توالی به دست می‌آید. از آنجا که بیان جزئیات بیشتر باعث دور شدن از مباحث اصلی می‌شود، در قسمت بعد تنها به آموزش برنامه جستجوی توالی‌ها در بانک‌های اطلاعاتی می‌پردازیم که بر مبنای هم‌ردیفی محلی دوگانه است.

جستجوی یک توالی

جستجوی بانک‌های اطلاعات توالی‌ها با یک توالی بر مبنای الگوریتم های نوشته شده برای هم ردیفی دوگانه صورت می‌گیرد.

اکنون، در بیشتر پایگاه‌ها از روشBLAST یا Basic Local Alignment Search Tools استفاده می‌شود.

[1] Gap extension penalty

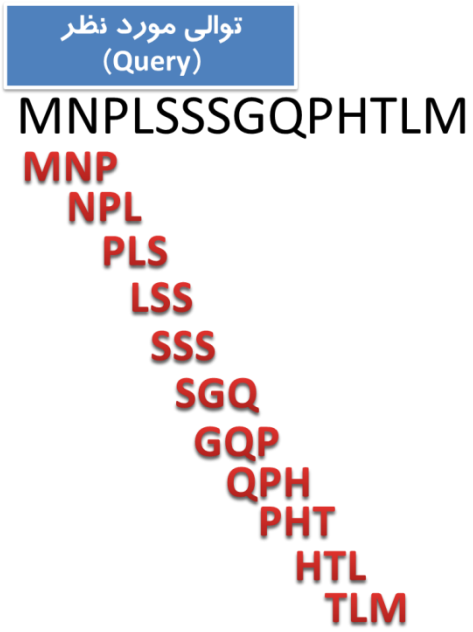
BLAST چگونه کار می‌کند

آنچه در برنامه BLAST انجام می شود پیدا کردن جفت قطعاتی مشابهی از توالی است که امتیاز هم‌ردیفی آنها از یک حد آستانه مشخصی بالاتر باشد. این قطعات HSPs (high-scoring segment pairs) نامیده می شوند. برای این کار برنامه BLAST از روش Dynamic Programming استفاده می‌کند.

در روشdynamic programming برای حل یک مشکل بزرگ، آن را به چند مشکل کوچک تجزیه می‌کنند. پس از یافتن پاسخ مناسب مشکلات کوچک، آنها را کنار هم چیده و راهی برای پاسخ به مشکل بزرگ پیدا می‌کنند. با توجه به طول بلند توالی‌ها و امکان جایگزینی و حذف و اضافه در آنها، جستجوی یک توالی در بین میلیون‌ها رکورد در بانک‌های اطلاعاتی نیازمند عملیات سنگینی است که از عهده ابررایانه‌های امروزی خارج است. در برنامه BLAST از روش dynamic programming برای حل این معضل استفاده شده است.

سه مرحله اصلی در الگوریتم BLAST وجود دارد که به شرح زیر هستند:

برنامه BLAST توالی مورد نظر (Query) را به قطعاتی با طول کوتاه یا کلمه (word) هم‌پوشان تبدیل می کند. معمولاً اندازهٔ کلمات برای توالی‌های آمینو اسیدی ۳ و برای توالی‌های نوکلئوتیدی ۱۱ تنظیم شده است. سپس از بین این کلمات آنهایی انتخاب می‌شوند که در یک هم‌ردیفی دوگانه با توالی الگو دارای امتیاز بالایی از یک حد تعیین شده هستند (مانند LSS). قابل ذکر است که امتیاز دهی بر اساس جداول PAM250 یا BLOSUM62 صورت می گیرد که توسط کاربر قابل تغییر است.



قطعات انتخاب شده برای جستجو در پایگاه توالی‌ها به کار رفته و توالی‌های مشابه ممکن (subjects) یافت می‌شوند.

برای هر کدام از جفت توالی‌های یافت شده با امتیاز بالا (HSP) هم‌ردیفی توالی از دو طرف کلمه ادامه پیدا می کند تا جایی که هم‌ردیفی جدیدی از امتیاز حد آستانه تعیین شده‌ای کمتر نشود. سپس اضافات توالیِ یافت شده حذف و هم‌ردیفی حاصل برای کاربر ارسال می‌شود.

انواع BLAST

بنا به نوع توالی مورد نظر و نوع پایگاه مورد جستجو، برنامه‌های BLAST طراحی شده‌اند که در زیر توضیح داده می‌شوند. البته برای هریک از این برنامه‌ها نیز زیربرنامه‌هایی که در آن تنظیمات، بهینه شده است معرفی شده‌اند، مانند BLASTN برای توالی‌های کوتاه.

- **BLASTN**

در این نوع BLAST، توالی مورد تقاضا نوکلئوتیدی است و جستجو در پایگاه توالی‌های نوکلئوتیدی انجام می‌شود. نتیجه جستجو جفت توالی‌های نوکلئوتیدی مشابه است که بر اساس شاخص‌های آماری میزان شباهت و یکسانی آنها نشان داده می‌شود.

- **BLASTP**

در این نوع BLAST، توالی مورد تقاضا، پروتئینی است و جستجو در پایگاه توالی‌های پروتئینی انجام می‌شود. نتیجهٔ جستجو، توالی‌های پروئینی مشابه با توالی الگو است که بر اساس شاخص‌های آماری، میزان شباهت و یکسانی آنها با توالی الگو نشان داده می‌شود.

- **BLASTX**

در این نوع BLAST، توالیِ مورد تقاضا نوکلئوتیدی است که در ۶ قالب خواندنی (ORF) ترجمه شده و به صورت توالی پروتئینی در پایگاه توالی‌های پروتئینی جستجو می‌شود. نتیجهٔ جستجو، توالی‌های مشابه با توالی الگو است که بر اساس آن می‌توانیم به توالی جدید خود قالب خواندنی و عملکرد نسبت بدهیم.

- **tBLASTN**

در این نوع BLAST، توالی مورد تقاضا پروتئینی است و جستجو در پایگاه توالی‌های نوکلئوتیدی انجام می شود که در ۶ قالب خواندنی ترجمه شده است . نتیجه جستجو توالی‌های مشابه با توالی مورد تقاضاست که بر اساس آن می‌توانیم برای توالی پروتئینی خود توالی‌های رمز کنندهٔ آن را شناسایی کنیم.

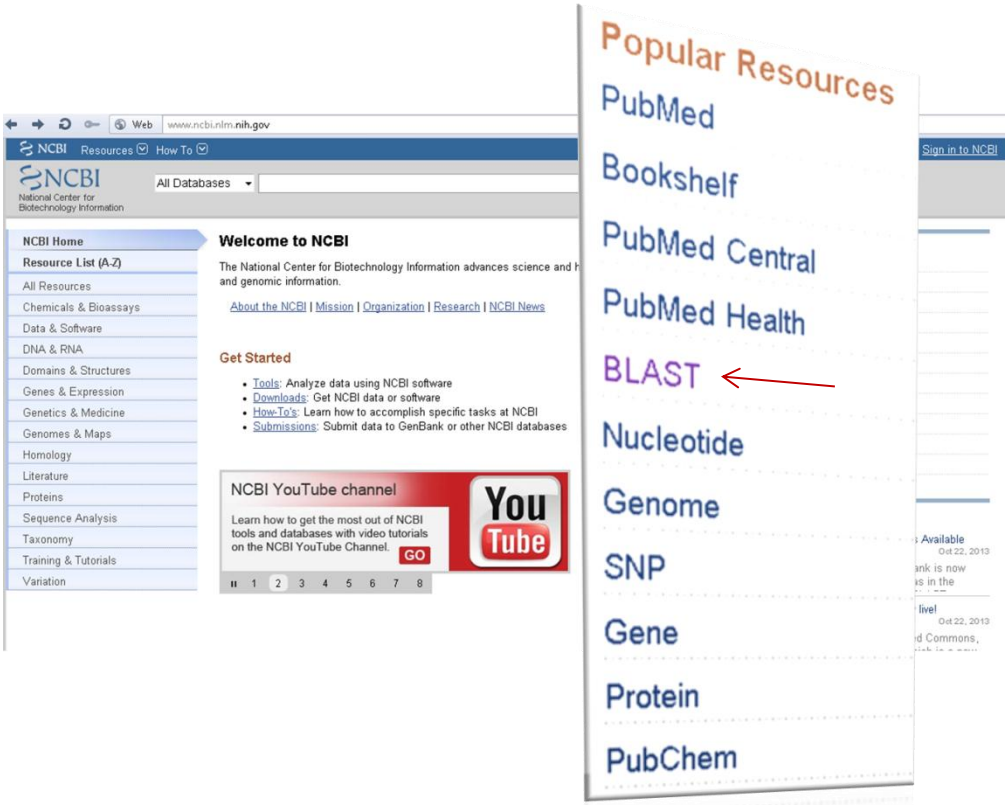
- **tBLASTX**

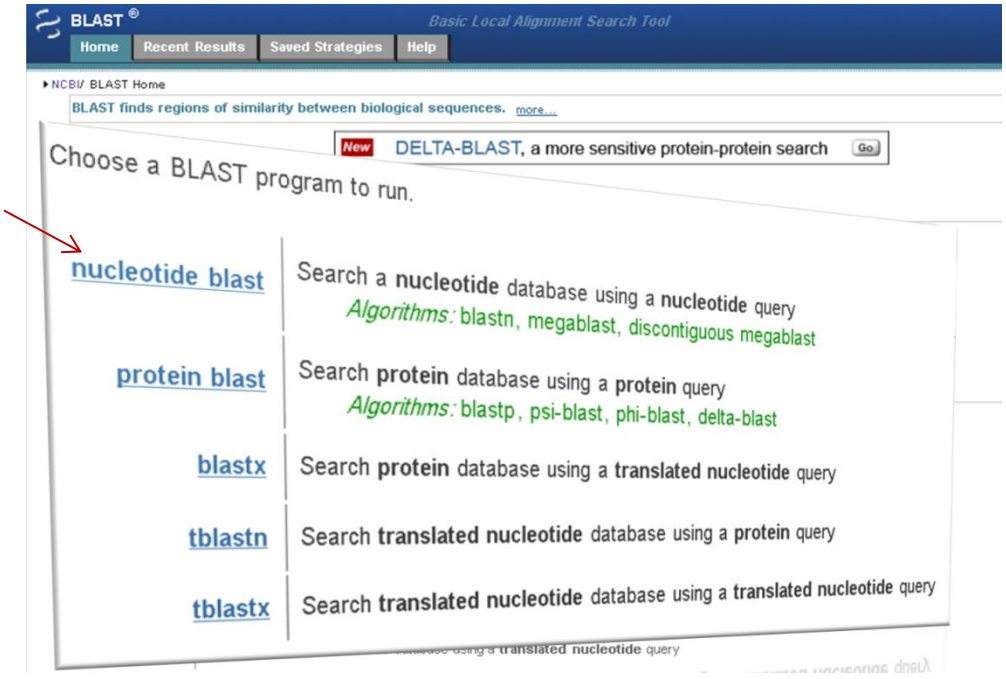
در این نوع BLAST، توالیِ مورد تقاضا، نوکلئوتیدی است که در ۶ قالب خواندنی به پروتئین ترجمه می‌شود ودر پایگاه توالی‌های نوکلئوتیدی که آن نیز در ۶ قالب خواندنی به پروتئین ترجمه می‌شود مورد جستجو قرار می گیرد. این نوع جستجو به ویژه در مطالعات EST به کار می رود.

وارد کردن توالی به منظور جستجو در صفحه BLAST

به طور معمول در تحقیقات، ما دارای یک توالی اولیه هستیم که می‌خواهیم توالی‌های مشابه آن را از طریق جستجو در بانک‌های اطلاعاتی به دست آوریم.

مراحل استفاده از **Blast** برای هم‌ردیف کردن یا **Align** دو توالی را در شکل‌های زیر می‌بینید. این شکل‌ها مراحل هم‌ردیف کردن توالی نوکلئوتیدی **Chitinase** از **Zebrafish** با توالی **Chitinase** از انسان را نشان می‌دهند.





Danio rerio chitinase, acidic.3 (chia.3), mRNA

NCBI Reference Sequence: NM_213213.1

GenBank Graphics

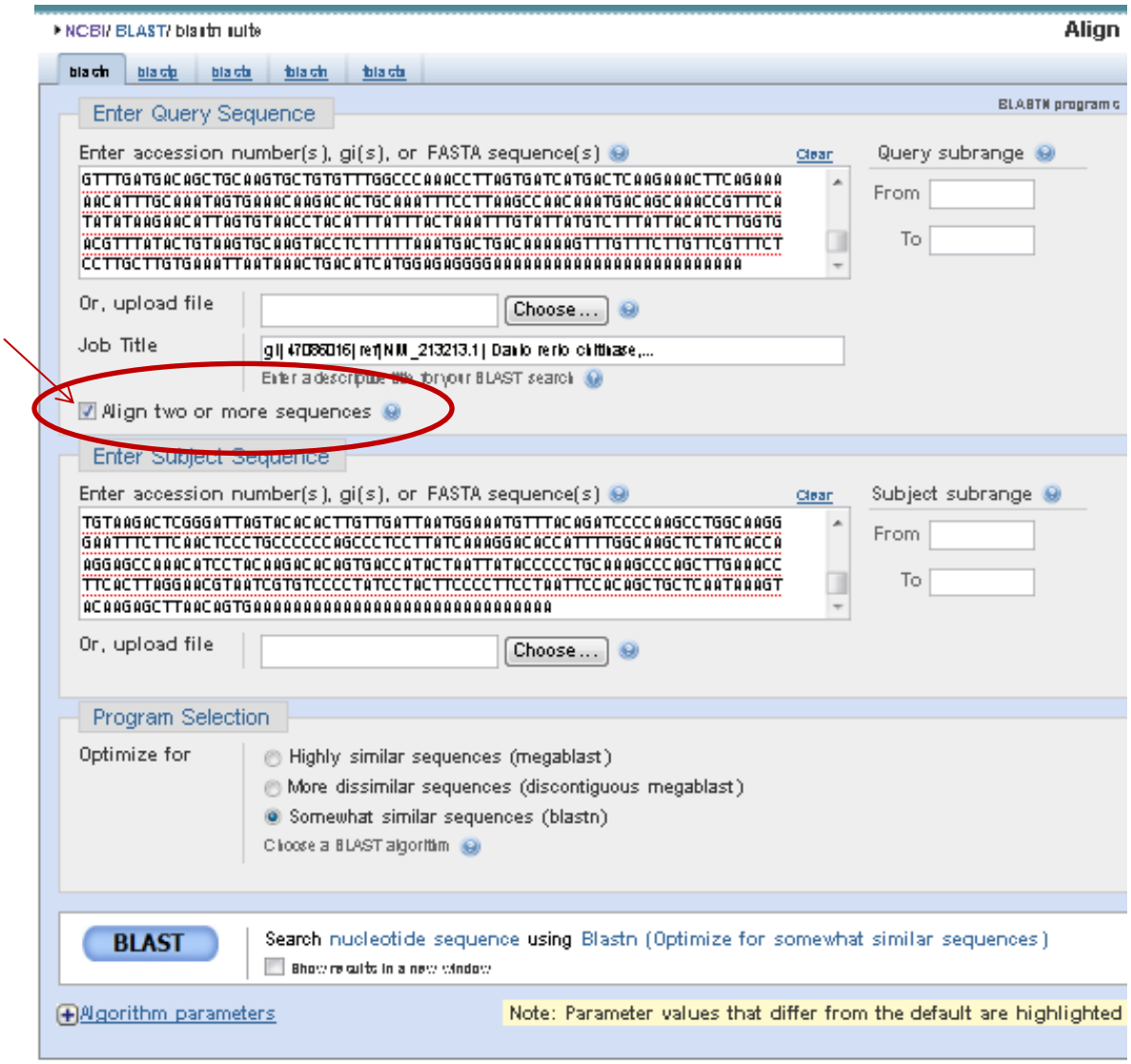
```
>gi|47086016|ref|NM_213213.1| Danio rerio chitinase, acidic.3 (chia.3), mRNA
GAGGTACCCCAGAAAATTGGCATTAGAATGGGGAGACTTACACTTATAGCAGGTTTGAGCTTGGTGCTCT
GCCATGTCGCCTTTTCCATGGAAATGGCCTGCTACTTCACCAACTGGTCCCAATATAGACCTGGAATTGG
AAAATATACTCCTGCAAATGTCGACCCTTACCTTTGCACACACCTTATCTATGCTTTTTCAATCATCAAC
CAAAGGAATGAGCTTGTCACATATG                    AGGCCTTCAATGAACTCAAGA
ACAAAAATCCCACTCTTAAGACCCT                    TGGTTCGGCACAGTTCTCCAT
CATGGTGTCCAATCCTGCAAACCGT                    AAATTCCTGAGAACTCATGGA
TTTGATGGACTGGATCTGGACTGGG                    CACCTGAAGACAAACAAAGAT
TCACTCTGCTGTGCAAGGAACTTGT                    AGCCACTGGCAATCCTCAGCT
TATGCTGACCGCTGCTGTATCAGCT                    TATGAGATTGCAGAGATTGCC
AAGTACTTGAACTTCATCAACGTCA                    GGGAGCGATTCACAGGACACA
ACAGCCCTCTGTACCAAGGCTCAAA                    CAACACCGACTATGCTATGCG
CTACTGGAGGGACAATGGAACCCCT                    GCAGCATACGGTCGCACTTTC
CGTCTGACATCTTCAGATACCAGCG                    CTTCAGCTGGAACCTACACTC
GCGAGGCTGGATTCTGGTCTTACTA                    AACAACAATTCAGTGGATTGA
TGACCAGAAGGTGCCCTATGCCACA                    GACACCAAGGAGAGTTATGAA
ACGAAGGTCCGTTATCTGAAAGACA                    GGGCACTTGATCTGGATGACT
TTGCTGGACAGTTCTGTAGTCAGGG                    TCGCAATCTTCTGGATATTGA
ATTGCCTCCAATGCCTTCAACTACC                    ACAAGGCCGACCACAACCACA
ACTACCACCACTCATGCTCCAGGACCAGGATTCTGTAATGGGAAGCCAGATGGACTCTATGCTCACCCTA
ATGACCCCAACAAATATTACAGCTGTGCTGGAGGTCATACCTTCGTGGAAAAATGTGCTGTAGGCACCGT
GTTTGATGACAGCTGCAAGTGCTGTGTTTGGCCCAAACCTTAGTGATCATGACTCAAGAAACTTCAGAAA
AACATTTGCAAATAGTGAAACAAGACACTGCAAATTTCCTTAAGCCAACAAATGACAGCAAACCGTTTCA
TATATAAGAACATTAGTGTAACCTACATTTATTTACTAAATTTGTATTATGTCTTTATTACATCTTGGTG
ACGTTTATACTGTAAGTGCAAGTACCTCTTTTTAAATGACTGACAAAAAGTTTGTTTCTTGTTCGTTTCT
CCTTGCTTGTGAAATTAATAAACTGACATCATGGAGAGGGGAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

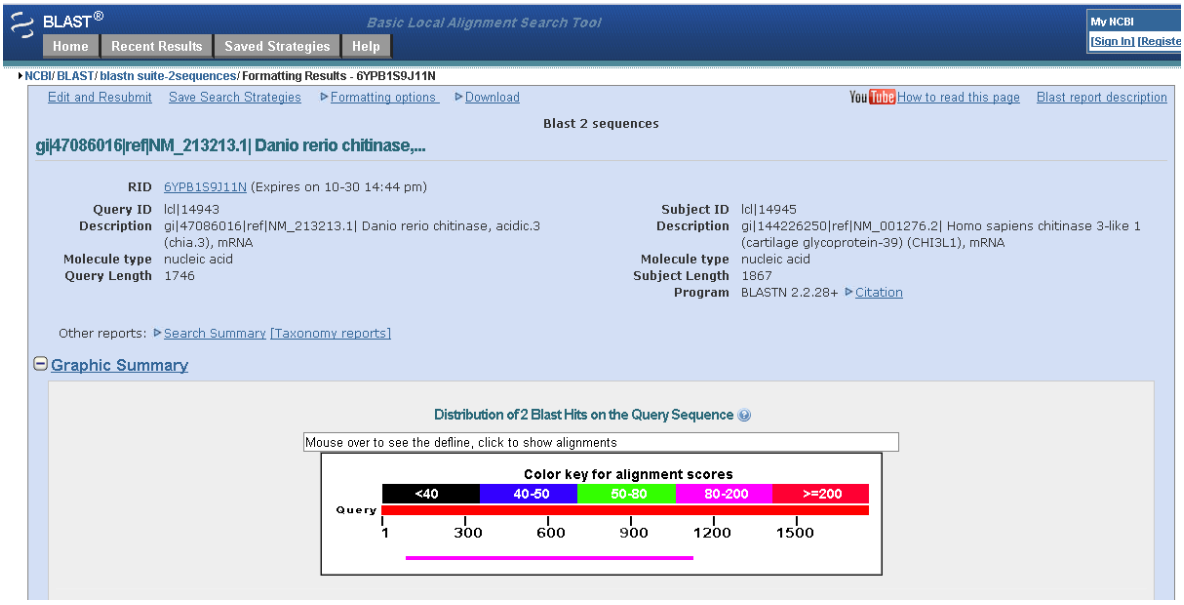

Homo sapiens chitinase 3-like 1 (cartilage glycoprotein-39) (CHI3L1), mRNA

NCBI Reference Sequence: NM_001276.2

GenBank Graphics

```
>gi|144226250|ref|NM_001276.2| Homo sapiens chitinase 3-like 1 (cartilage glycoprotein-39)
(CHI3L1), mRNA
CACATAGCTCAGTTCCCATAAAAGGGCTGGTTTGCCGCGTCGGGGAGTGGAGTGGGACAGGTATATAAAG
GAAGTACAGGGCCTGGGGAAGAGGCCCTGTCTAGGTAGCTGGCACCAGGAGCCGTGGGCAAGGGAAGAGG
CCACACCCTGCCCTGCTCTGCTGCAGCCAGAATGGGTGTGAAGGCGTCTCAAACAGGCTTTGTGGTCCTG
GTGCTGCTCCAGTGCTGCTCTGCATACAAACTGGTCTGCTACTACACCAGCTGGTCCCAGTACCGGGAAG
GCGATGGGAGCTGCTTCCCAGATGCCCTTGACCGCTTCCTCTGTACCCACATCATCTACAGCTTTGCCAA
TATAAGCAACGATCACATCGACACCTGGGAGTGGAATGATGTGACGCTC
AAGAACAGGAACCCCAACCTGAAGACTCTCTTGTCTGTCGGAGGATGGA
CCAAGATAGCCTCCAACACCCAGAGTCGCCGGACTTTCATCAAGTCAGT
TGGCTTTGATGGGCTGGACCTTGCCTGGCTCTACCCTGGACGGAGAGAC
ATCAAGGAAATGAAGGCCGAATTTATAAAGGAAGCCCAGCCAGGGAAAA
CACTGTCTGCGGGGAAGGTCACCATTGACAGCAGCTATGACATTGCCAA
CATTAGCATCATGACCTACGATTTTCATGGAGCCTGGCGTGGGACCACA
CGAGGTCAGGAGGATGCAAGTCCTGACAGATTCAGCAACACTGACTATG
TGGGGGCTCCTGCCAGTAAGCTGGTGATGGGCATCCCCACCTTCGGGAG
TGAGACTGGTGTTGGAGCCCCAATCTCAGGACCGGGAATTCCAGGCCGG
CTTGCCTACTATGAGATCTGTGACTTCCTCCGCGGAGCCACAGTCCATA
CCTATGCCACCAAGGGCAACCAGTGGGTAGGATACGACGACCAGGAAAG
CCTGAAGGACAGGCAGCTGGCGGGCGCCATGGTATGGGCCCTGGACCTG
TGCGGCCAGGATCTGCGCTTCCCTCTCACCAATGCCATCAAGGATGCAC
TTCTGCACACAGCACGGGGGCCAAGGATGCCCCGTCCCCCTCTGGCTCC
CCTGCCCTGCTGAGTCCCAGGCTGAGCCTCAGTCTCCCTCCCTTGGGGCCTATGCAGAGGTCCACAACAC
ACAGATTTGAGCTCAGCCCTGGTGGGCAGAGAGGTAGGGATGGGGCTGTGGGGATAGTGAGGCATCGCAA
TGTAAGACTCGGGATTAGTACACACTTGTTGATTAATGGAAATGTTTACAGATCCCCAAGCCTGGCAAGG
GAATTTCTTCAACTCCCTGCCCCCCAGCCCTCCTTATCAAAGGACACCATTTTGGCAAGCTCTATCACCA
AGGAGCCAAACATCCTACAAGACACAGTGACCATACTAATTATACCCCCTGCAAAGCCCAGCTTGAAACC
TTCACTTAGGAACGTAATCGTGTCCCCTATCCTACTTCCCCTTCCTAATTCCACAGCTGCTCAATAAAGT
ACAAGAGCTTAACAGTGAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```


NCBI/ BLAST/ blastn suite
Align
blastn
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Choose...
Job Title
Align two or more sequences
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Subject subrange
From
To
Or, upload file
Choose...
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted

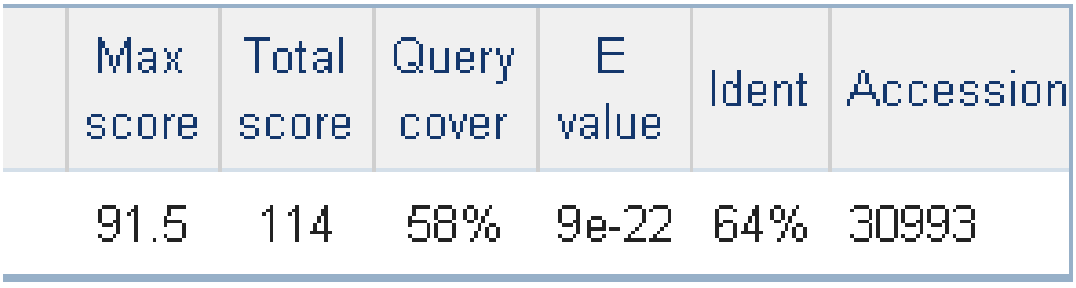


Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
gi|144226250|ref|NM_001276.2| Homo sapiens chitinase 3-like 1 (cartilage glycoprotein-39) (CHI3L1), mRNA	91.5	114	58%	9e-22	64%	30993



gi|144226250|ref|NM_001276.2| Homo sapiens chitinase 3-like 1 (cartilage glycoprotein-39) (CHI3L1), mRNA

Sequence ID: lcl|30993 **Length:** 1867 **Number of Matches:** 2

Range 1: 239 to 1247 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
91.5 bits(100)	9e-22	669/1051(64%)	63/1051(5%)	Plus/Plus

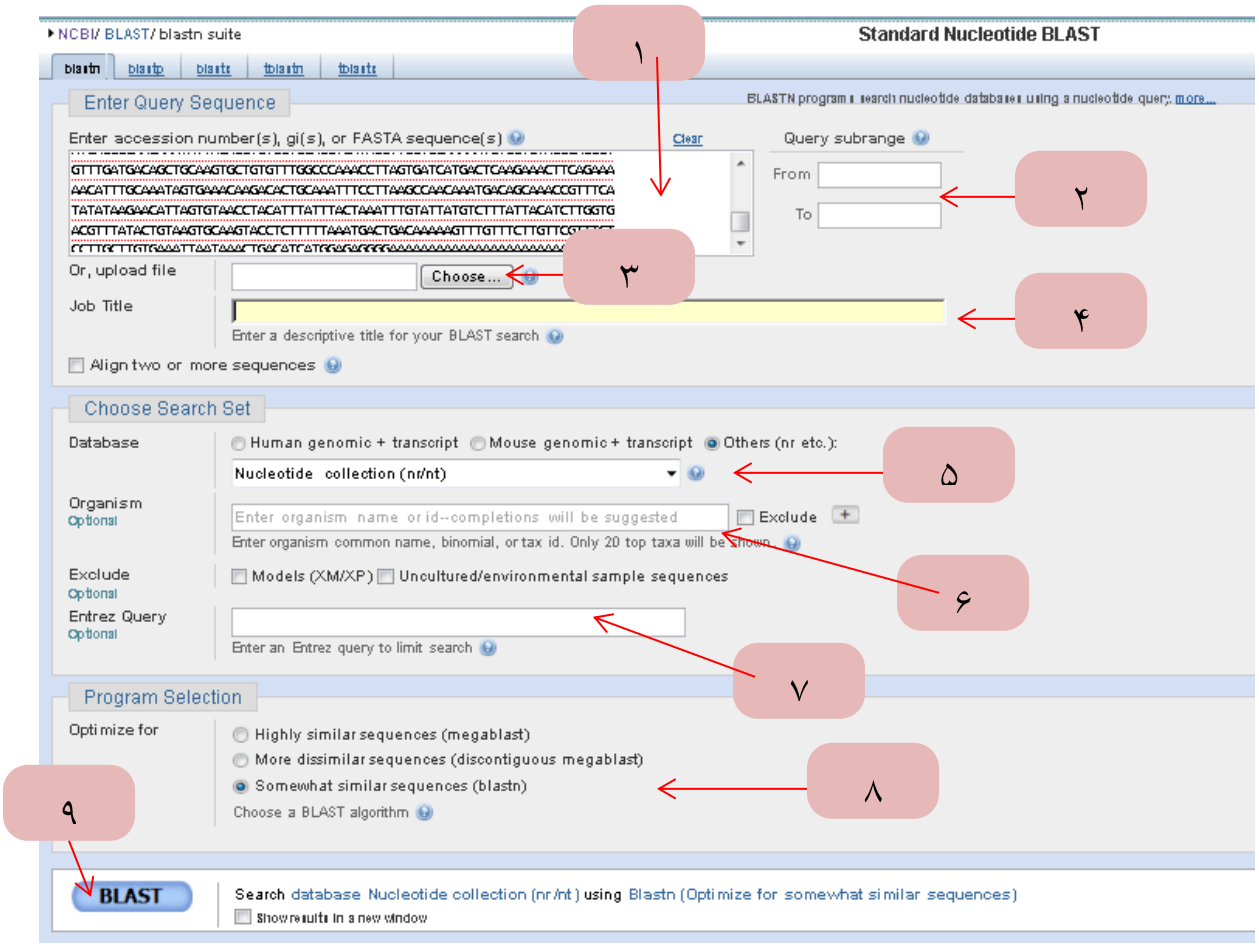
```
Query  92    AAATGGCCTGCTACTTCACCAACTGGTCCCAATATAGACCTGGAATTGGAAAATATACTC  151
             || ||| ||||||| ||||| ||||||||| ||    || ||||  ||    |  ||
Sbjct  239   AACTGGTCTGCTACTACACCAGCTGGTCCCAGTA----CCGGGAAGGCGATGGGAG-CTG  293

Query  152   CTGC--AAATGTC---GACCCTTACCTTTGCACACACCTTATCTAT-GCTTTTTCAATCA  205
             || |  | ||| |   ||||  | ||| || || ||| | |||||  |||||  |||| |
Sbjct  294   CTTCCCAGATGCCCTTGACCGCTTCCTCTGTACCCACATCATCTACAGCTTTGCCAAT-A  352

Query  206   TCAACCAAAGGAATGAGCTTGTCACATATGAGTGGAACGATGAAACTCTATACAAGGCCT  265
             | || ||| |  || |  | | ||| |  |||||||| ||||  || || |||  ||| |
Sbjct  353   T-AAGCAACG--ATCACATCGACACCTGGGAGTGGAATGATGTGACGCTCTAC--GGCAT  407

Query  266   --TCAATGAACTCAAGAACAAAAATCCCACTCTTAAGACCCTTTTGGCTGTTGGAGGATG  323
               ||||   ||||||||||  || ||||  || ||||| || ||| |||| ||||||||
Sbjct  408   GCTCAACACACTCAAGAACAGGAACCCCAACCTGAAGACTCTCTTGTCTGTCGGAGGATG  467

Query  324   GAATTTTGGTTCGGCACAG-TTCTCCATCATGGTGTCCAATCCTGCAAACCGTAAAACAT  382
             ||| ||||| ||  || || || ||||  || |  |||||  |    |  ||    || |
Sbjct  468   GAACTTTGGGTCT-CAAAGATTTTCCAAGATAGCCTCCAACACCCAGAGTCGCCGGACTT  526

Query  383   TTATTCAGTC--TACCATCAAATTCCTGAGAACTCATGGATTTGATGGACTGGATCTGGA  440
             | ||  ||||  ||||  ||  || ||| | || ||||| |||||||| ||||| || |
Sbjct  527   TCATCAAGTCAGTACCGCCA--TTTCTGCGCACCCATGGCTTTGATGGGCTGGACCTTGC  584

Query  441   CTGGGAATATCCCGGAGCAAGAGGAAGTCCACCTGAAGACAAACAAAGATTCACTCTGCT  500
             ||||   || || |||   ||||                |||||||    || ||    ||
Sbjct  585   CTGGCTCTACCCTGGACGGAGAG---------------ACAAACAGCATTTTACCACCCT  629

Query  501   GTGCAAGGAACTTGTTGCAGCCT--ATGAGGCTGAGAGTAAAGCCACTGGCAATCCTCAG  558
                |||||| |    || |  |  || | |  || ||   |||||  || ||     |||
Sbjct  630   AATCAAGGAAATGAAGGCCGAATTTATAAAG--GA-AGCCCAGCCA--GGGAAAAAGCAG  684

Query  559   CTTATGCTGACCGCTGCTGTATCAGCTGGCAAAGGCACTATTGATGATGGATATGAGATT  618
             ||  |||| | ||| ||  | || || || || | ||| |||||     | ||||| |||
Sbjct  685   CTCCTGCTCAGCGCAGCACTGTCTGCGGGGAAGGTCACCATTGACAGCAGCTATGACATT  744

Query  619   GCAGAGATTGCCAAGTACTTGAACTTCATCAACGTCATGACTTACGACTTCCATGGCACT  678
             ||  ||||  || |  || || | ||||| | | ||||||| ||||| || |||||  |
Sbjct  745   GCCAAGATATCCCAACACCTGGATTTCATTAGCATCATGACCTACGATTTTCATGGAGCC  804

Query  679   TGG-GAGCGATTCACAGGACACAACAGCCCTCTGTACCAAGGCTCAA-AGGATGAGGGAG  736
             ||| | | ||  |||||| ||  |||| || |||| || ||| |||  ||||||    ||
Sbjct  805   TGGCGTGGGAC-CACAGGCCATCACAGTCCCCTGTTCCGAGG-TCAGGAGGATGCA--AG  860

Query  737   ACCTGATCTACTT--CAACACCGACTATGCTATGCGCTAC-TGGAGGGACAATGGAACCC  793
               |||||   | |   |||||| ||||||||| || | ||| ||  | | |   ||  | |
Sbjct  861   TCCTGACAGATTCAGCAACACTGACTATGCTGTGGGGTACATGTTGAGGCTGGGGG-CTC  919

Query  794   CTGTGGAGAAACTCAGAATGGGCTTTGCAGCATACGGTCGCACTTTCCGTCTGACATCTT  853
             |||     || ||    |||||| |  |  | | |||  | |  |||  |||| | ||||
Sbjct  920   CTGCCAGTAAGCTGGTGATGGGCATCCCCACCTTCGGGAGGAGCTTCACTCTGGCTTCTT  979

Query  854   CAGATACCAGCGTTGGAGCTCCAGCTAGTGGACCTGC--TTCAGCTGGAACCTACACTCG  911
             | || ||  | |||||||| |||      ||||| |   |||  | ||     | |||
Sbjct  980   CTGAGACTGGTGTTGGAGCCCCAATCTCAGGACCGGGAATTC--CAGGCCGGTTCACCAA  1037

Query  912   CGAGGCTGG-ATTCTGGTCTTACTATGAGATCTGTGGATTCCTGGAGGGAACAACAATTC  970
              ||||| || |  || | || ||||||||||||||  |||||    ||| | ||| | |
Sbjct  1038  GGAGGCAGGGACCCTTGCCT-ACTATGAGATCTGTGACTTCCTCCGCGGAGCCACAGTCC  1096

Query  971   AGTGGATTGATGACCAGAAGGTGCCCTATGCCACAAAGAACAGCGAGTGGGTTGGATTTG  1030
             |  | ||     | |||| |||| ||||||||||| |||  || | |||||| ||||  |
Sbjct  1097  ATAGAATCCTCGGCCAGCAGGTCCCCTATGCCACCAAGGGCAACCAGTGGGTAGGATACG  1156

Query  1031  ACACCAAGGAGAGTTATGAAACGAAGGTCCGTTATCTGAAAGACAAGAATTTTGGTGGAG  1090
             ||  | |||| ||     |||  ||||| |  || ||||| |||| | |  | |  || |
Sbjct  1157  ACGACCAGGAAAGCGTCAAAAGCAAGGTGCAGTACCTGAAGGACAGGCAGCTGGCGGGCG  1216

Query  1091  CTTTCGTTTGGGCACTTGATCTGGATGACTT  1121
             |  | || ||||| || || ||||||||||||
Sbjct  1217  CCATGGTATGGGCCCTGGACCTGGATGACTT  1247
```

Range 2: 1729 to 1740 Graphics ▼ Next Match ▲ Previous Match ▲ Firs

Score	Expect	Identities	Gaps	Strand
22.9 bits(24)	0.40	12/12(100%)	0/12(0%)	Plus/Minus

```
Query  812   TGGGCTTTGCAG  823
             ||||||||||||
Sbjct  1740  TGGGCTTTGCAG  1729
```

مراحل استفاده از BLAST برای مقایسه یک توالی با سایر توالی‌های موجود در بانک اطلاعاتی



۱. در این قسمت توالی مورد نظر را بافرمت FASTA یا شمارهٔ دسترسی آن را در صورت دانستن وارد می‌نماییم.

۲. در این قسمت می‌توانیم قسمتی از توالی را مشخص کنیم که می‌خواهیم مورد جستجو قرار گیرد.

۳. با استفاده از این بخش می‌توانیم توالی را از جایی که ذخیره کرده‌ایم upload کنیم.

۴. در این بخش می‌توانیم برای جستجوی خود توضیح مختصری وارد کنیم.

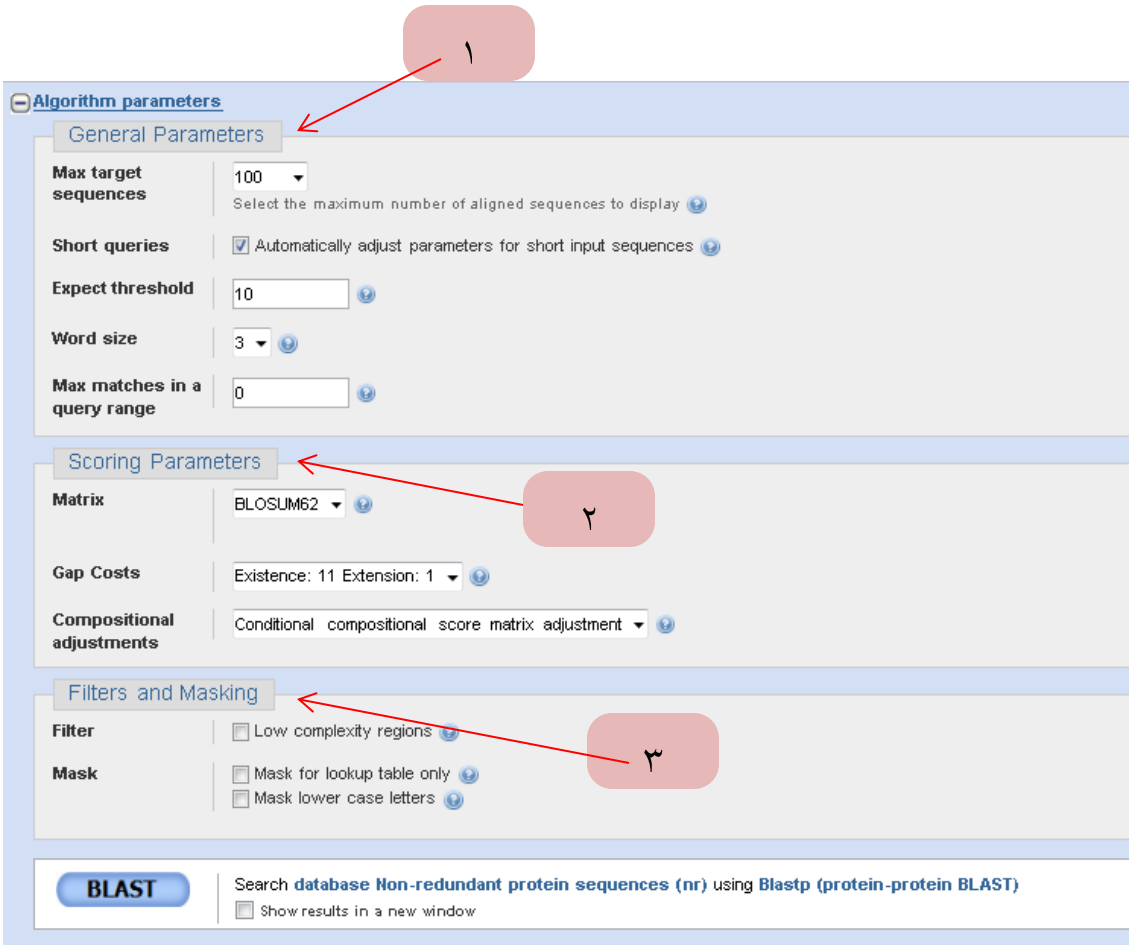
۵. این بخش مربوط به نوع پایگاه داده است که می خواهیم توالی ما در آن جستجو شود. به صورت پیش فرض کلیه پایگاه‌های مشابه یا nr (nonredundent) مورد هدف است.

۶. این بخش مربوط به انتخاب نوع موجود زنده مورد نظر ماست که گزینه‌ای اختیاری است و در صورتی از آن استفاده می‌کنیم که بخواهیم فقط در موجود خاصی توالی ما جستجو شود.

۷. این بخش مربوط به جستجوی توالی ما در ENTREZ است که اختیاری است و می‌توانیم جستجو در مورد توالی خود را با کلید واژه محدود کنید.

۸. در این بخش نوع الگوریتم BLAST را مشخص می‌کنیم.

۹. با کلیک بر روی نمایهٔ BLAST جستجو آغاز می‌شود.



۱. این بخش مربوط به تنظیم پارامترهای الگوریتم‌های مورد استفاده در BLAST است که شامل پارامترهای عمومی، امتیازدهی و فیلترگذاری می‌شود. در این بخش می‌توان تعداد توالی‌های نمایشی در یک صفحه، پارامترهای مربوط به توالی‌های کوتاه، میزان آستانه و اندازهٔ کلمات مورد جستجو توسط BLAST را تنظیم کرد.

۲. این قسمت برای تنظیم پارامترهای امتیازدهی طراحی شده است. نوع ماتریس امتیازدهی، جریمه فواصل و فرض‌های مربوط به اعمال ماتریس‌های امتیازی را می‌توان به صورت دستی تنظیم کرد.

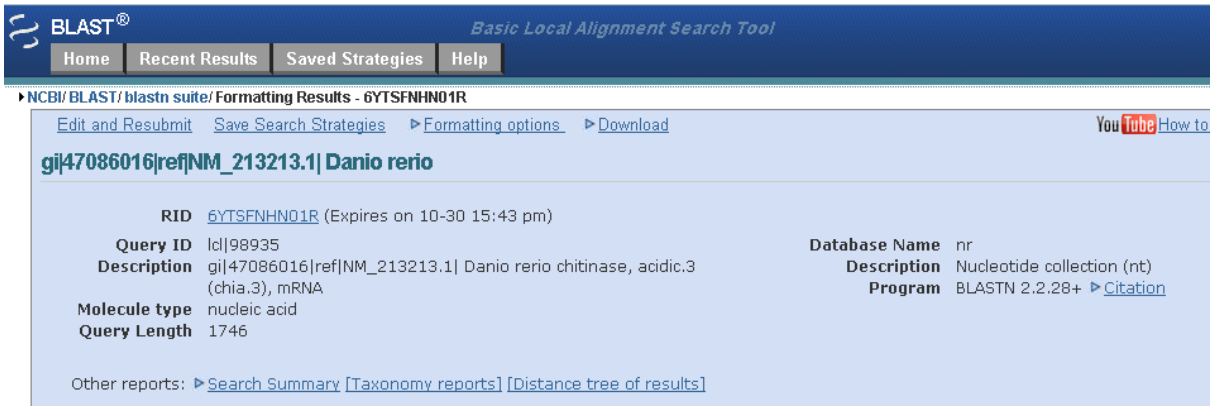
۳. در این بخش می‌توان الگوریتم را برای در نظر نگرفتن نواحی از توالی که ارزش تکاملی ندارند و یا دارای اشکال و ابهام هستند تنظیم کرد. با انتخاب فیلتر مربوطه بخش‌هاط موردنظر از توالی در جستجو لحاظ نمی‌شود.

صفحه نتایج

بعد از وارد کردن توالی در جایگاه مربوطه و انتخاب پارامترهای مورد نظر، زمانی که کلید BLAST را می‌زنیم صفحه‌ای باز می‌شود که اطلاعات مختصری را در مورد توالی ما می‌دهد و اعلام می‌کند که جستجوی ما در حال انجام است و نتایج، بعد از چند ثانیه نشان داده خواهند شد. در صورت شلوغ بودن، سرور سایت اعلام می‌کند که جستجوی خود را در زمانی دیگر دوباره انجام دهیم.

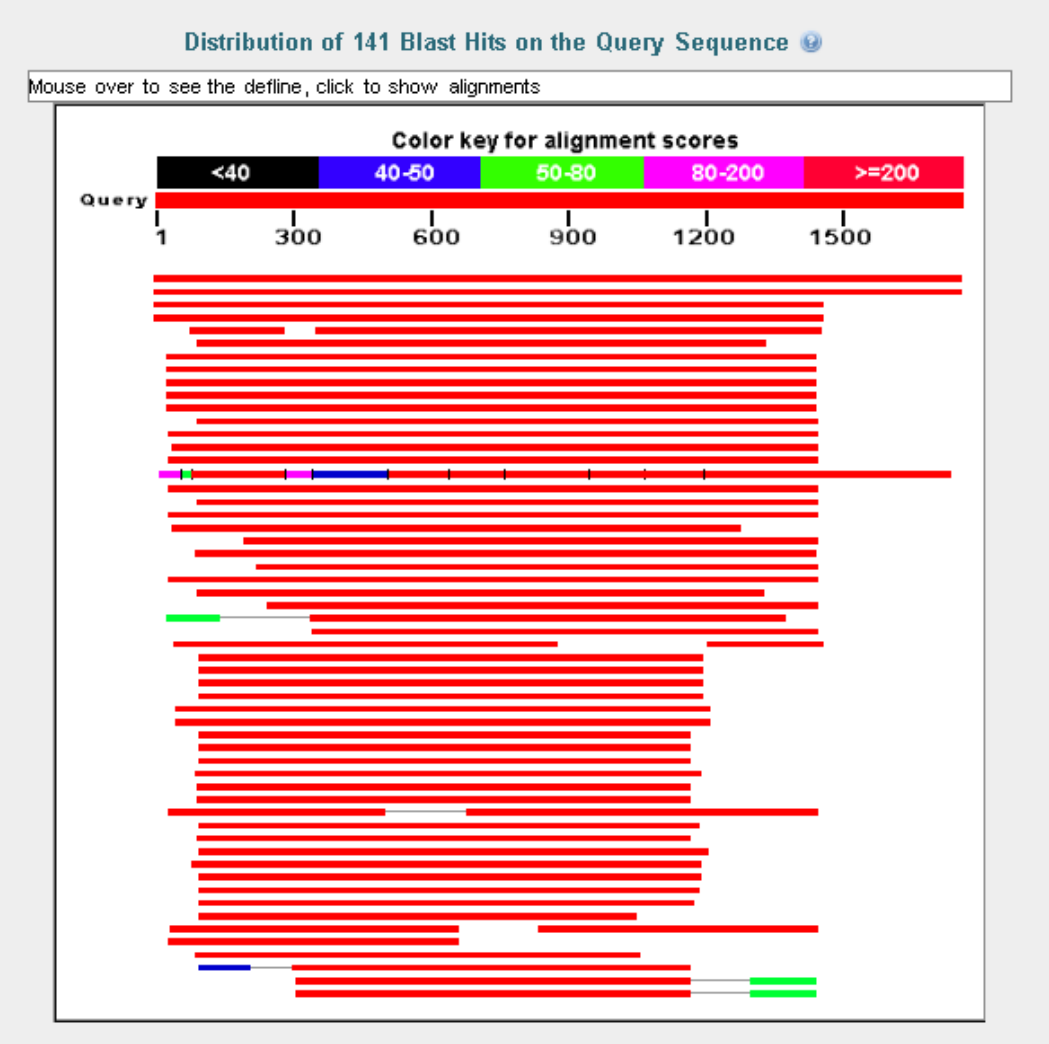
صفحهٔ نتایج شامل سه بخش است:

بخش اول شامل توضیحاتی در مورد در مورد الگوریتم BLAST و نویسندگان این الگوریتم و ویرایش هایی است که در گذشته و حال مورد استفاده هستند. در این بخش جستجوی ما دارای یک ID است که در صورت لزوم (مثلا بروز مشکل و درخواست کمک از مدیر پایگاه) می توانیم با این ID به جستجوی خود در NCBI دسترسی داشته باشیم. این بخش توضیحاتی در مورد پایگاه داده‌ای که ما برای جستجوی خود انتخاب کرده‌ایم نیز می دهد.



بخش اول نتایج Blast

بخش دوم این صفحه نمایشی گرافیکی از نتایج BLAST است به طوری که ۱۰۰ توالی اولی که در جستجوی BLAST به دست آمده‌اند به صورت خطوط رنگی نشان داده می‌شوند، هر توالی بر اساس میزان شباهت خود دارای یک طیف رنگی است .کلید رمز رنگ‌ها بر اساس امتیاز هم‌ردیفی در بالای آن آمده است. هر چه قدر جور شدن توالی یافت شده با توالیِ در حال جستجو بیشتر باشد با رنگ قرمز و هر چه کمتر باشد با رنگ‌های رو به مشکی نمایش داده می‌شود. بنابراین در این قسمت در یک نگاه کلی می‌توانیم میزان شباهت را مشاهده کنیم.



بخش دوم نتایج Blast

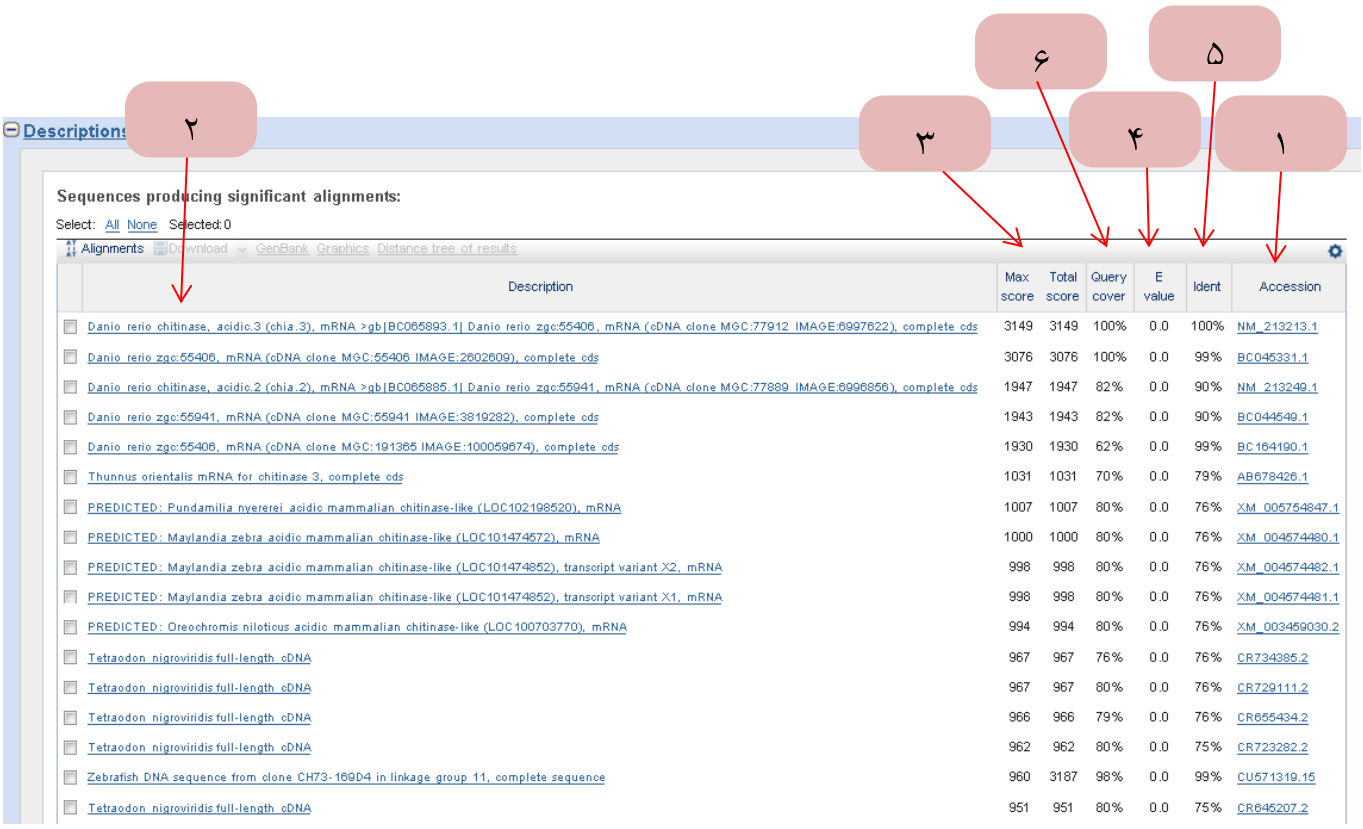
بخش سوم

در بخش سوم شماره دسترسی (شماره ۱) و نام (شماره ۲) توالی‌های به دست آمده فهرست شده‌اند. روبروی هر توالی اعدادی نوشته شده است. اولین عدد امتیاز همردیفی دوگانه (شماره ۳) توالی یافت شده و توالی در حال جستجو است که کلیه اطلاعات بعدی بر اساس آنها مرتب می‌شوند و شامل Max Score (بیشترین امتیاز مربوط به یک قطعه از Subject) و Total Score (امتیاز کل یا مجموع امتیازات قطعات هم‌ردیف‌شده از یک Subject). دومین شاخص، ارزش مورد انتظار یا E-value (شماره ۴) است. تعریف این شاخص به صورت ساده این است:

- چه قدر احتمال دارد که توالی جفت شده با توالی الگوی ما به طور تصادفی جفت شده باشد و هیچ رابطهٔ معنی‌داری بین آنها وجود نداشته باشد.

- طبیعی است که اگر E-Value صفر باشد، ایده‌آل‌ترین حالت ممکن است و در غیر این صورت هر چه به صفر نزدیک‌تر باشد اطمینان ما به نتیجهٔ به‌دست آمده بیشتر می شود. نتایج BLAST بر اساس این دو شاخص مرتب می‌شوند و بنابراین توالی‌هایی که در اوائل لیست هستند توالی هایی هستند که اطمینان ما در شباهت آنها به توالی الگوی مورد نظرمان بیشتر است.

شماره ۵ درصد همسانی (Identity) و شماره ۶، Query cover یا درصدی از Query را که توسط Subject پوشش داده شده است، نشان می‌دهد.

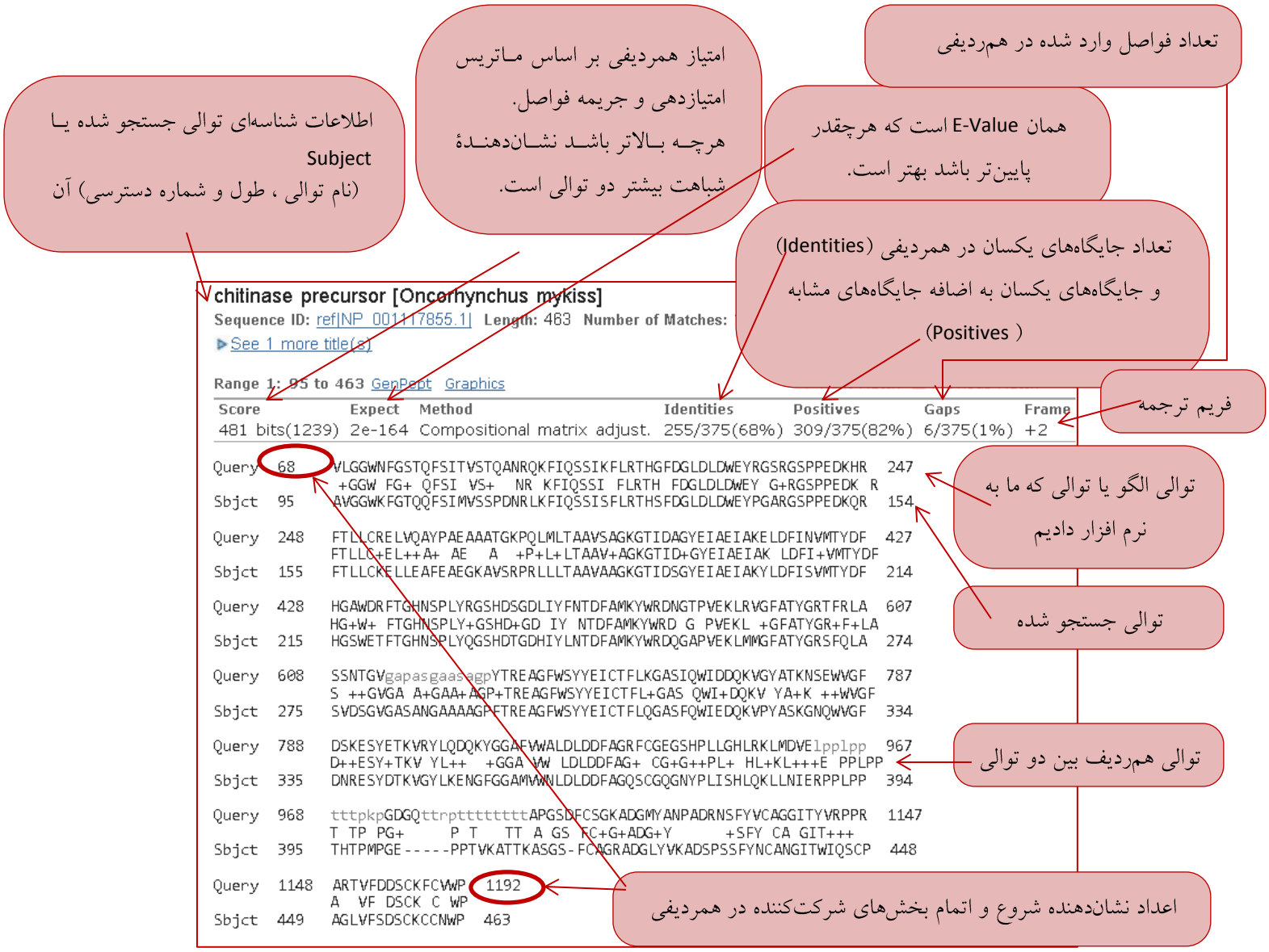


Description	Max score	Total score	Query cover	E value	Ident	Accession
Danio rerio chitinase, acidic.3 (chia.3), mRNA >gb|BC065893.1| Danio rerio zgc:55406, mRNA (cDNA clone MGC:77912 IMAGE:6997622), complete cds	3149	3149	100%	0.0	100%	NM_213213.1
Danio rerio zgc:55406, mRNA (cDNA clone MGC:55406 IMAGE:2602609), complete cds	3076	3076	100%	0.0	99%	BC045331.1
Danio rerio chitinase, acidic.2 (chia.2), mRNA >gb|BC065885.1| Danio rerio zgc:55941, mRNA (cDNA clone MGC:77889 IMAGE:6996856), complete cds	1947	1947	82%	0.0	90%	NM_213249.1
Danio rerio zgc:55941, mRNA (cDNA clone MGC:55941 IMAGE:3819282), complete cds	1943	1943	82%	0.0	90%	BC044549.1
Danio rerio zgc:55406, mRNA (cDNA clone MGC:191365 IMAGE:100059674), complete cds	1930	1930	62%	0.0	99%	BC164190.1
Thunnus orientalis mRNA for chitinase 3, complete cds	1031	1031	70%	0.0	79%	AB678426.1
PREDICTED: Pundamilia nyererei acidic mammalian chitinase-like (LOC102198520), mRNA	1007	1007	80%	0.0	76%	XM_005754847.1
PREDICTED: Maylandia zebra acidic mammalian chitinase-like (LOC101474572), mRNA	1000	1000	80%	0.0	76%	XM_004574480.1
PREDICTED: Maylandia zebra acidic mammalian chitinase-like (LOC101474852), transcript variant X2, mRNA	998	998	80%	0.0	76%	XM_004574482.1
PREDICTED: Maylandia zebra acidic mammalian chitinase-like (LOC101474852), transcript variant X1, mRNA	998	998	80%	0.0	76%	XM_004574481.1
PREDICTED: Oreochromis niloticus acidic mammalian chitinase-like (LOC100703770), mRNA	994	994	80%	0.0	76%	XM_003459030.2
Tetraodon nigroviridis full-length cDNA	967	967	76%	0.0	76%	CR734385.2
Tetraodon nigroviridis full-length cDNA	967	967	80%	0.0	76%	CR729111.2
Tetraodon nigroviridis full-length cDNA	966	966	79%	0.0	76%	CR655434.2
Tetraodon nigroviridis full-length cDNA	962	962	80%	0.0	75%	CR723282.2
Zebrafish DNA sequence from clone CH73-169D4 in linkage group 11, complete sequence	960	3187	98%	0.0	99%	CU571319.15
Tetraodon nigroviridis full-length cDNA	951	951	80%	0.0	75%	CR645207.2

بخش سوم نتایج Blast.

در بخش چهارم جزئیات هم‌ریفیِ تک توالی‌ها با توالی الگوی ما آورده شده است که شامل اطلاعاتی در مورد Score هم‌ردیفی، میزان شباهت دو توالی و تعداد جایگاه‌های جفت‌شده و تعداد فواصل استفاده شده و ... است. ضمن اینکه بخشی از توالی که هم‌ردیف شده‌اند در این توضیحات می‌آید که می‌توان به راحتی مناطق مشترک بین دو توالی را مشاهده کرد.

یکی از هم‌ردیفی‌ها، نتیجه BLASTX:



بخش چهارم نتایج Blast.

نمونه‌ای از نتایج Blastn

PREDICTED: Oreochromis niloticus acidic mammalian chitinase-like (LOC100703770), mRNA

Sequence ID: ref|XM_003459030.2| **Length:** 1580 **Number of Matches:** 1

Range 1: 89 to 1506 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
994 bits(1102)	0.0	1078/1424(76%)	24/1424(1%)	Plus/Plus

```
Query  28    ATGGGGAGACTTACACTTATAGCAGGTTTGAGCTTGGTGCTCTGCCATGTCGCCTTTTCC  87
             ||||  ||  ||  ||| ||  ||||||||||| ||  ||||| |  ||||  | |  | | ||
Sbjct  89    ATGGCCAGGCTCACAATTCTAGCAGGTTTGTGCCTGGTGATAAGCCAGCTGGGATCTGCC  148

Query  88    ATGGAAATGGCCTGCTACTTCACCAACTGGTCCCAATATAGACCTGGAATTGGAAAATAT  147
             |     ||||  ||||||||||||||||||||||| ||  ||  ||||||  |||||| |
Sbjct  149   AGCAGGATGGAGTGCTACTTCACCAACTGGTCCCAGTACAGGCCTGGAGATGGAAAGT--  206
```

دو توالی به صورت هم جهت با یکدیگر هم‌ردیف شده‌اند یا به عبارت دیگر هر دو روی رشته بالا (Sense) قرار گرفته‌اند

Range 5: 4543 to 4677 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Gaps	Strand
244 bits(270)	2e-60	135/135(100%)	0/135(0%)	Plus/Minus

```
Query  1055  AGGTCCGTTATCTGAAAGACAAGAATTTTGGTGGAGCTTTCGTTTGGGCACTTGATCTGG  1114
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4677  AGGTCCGTTATCTGAAAGACAAGAATTTTGGTGGAGCTTTCGTTTGGGCACTTGATCTGG  461

Query  1115  ATGACTTTGCTGGACAGTTCTGTAGTCAGGGGAACCATCCTCTCATGGCCCATCTTCGCA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4617  ATGACTTTGCTGGACAGTTCTGTAGTCAGGGGAACCATCCTCTCATGGCCCATCTTCGCA

Query  1175  ATCTTCTGGATATTG  1189
              |||||||||||||||
Sbjct  4557  ATCTTCTGGATATTG  4543
```

دو توالی به صورت غیر هم‌جهت با یکدیگر هم‌ردیف شده‌اند یا به عبارت دیگر یکی روی رشته بالا (Sense) و دیگری روی رشته پایین (anti-sense) قرار گرفته است.

منبع:

ملبوبی، م.ع.، فیضی، ا.، لهراسبی، ت. راهنمای عملی داده‌پردازی زیستی و پروژه‌های ژنوم. انتشارات استاد ملبوبی.