

بسمه تعالی

بیوانفورماتیک

گروه بیوتکنولوژی دریا

کارشناسی ارشد



معرفی پایگاه‌های اطلاعاتی اولیه و ابزارهای جستجو در آنها

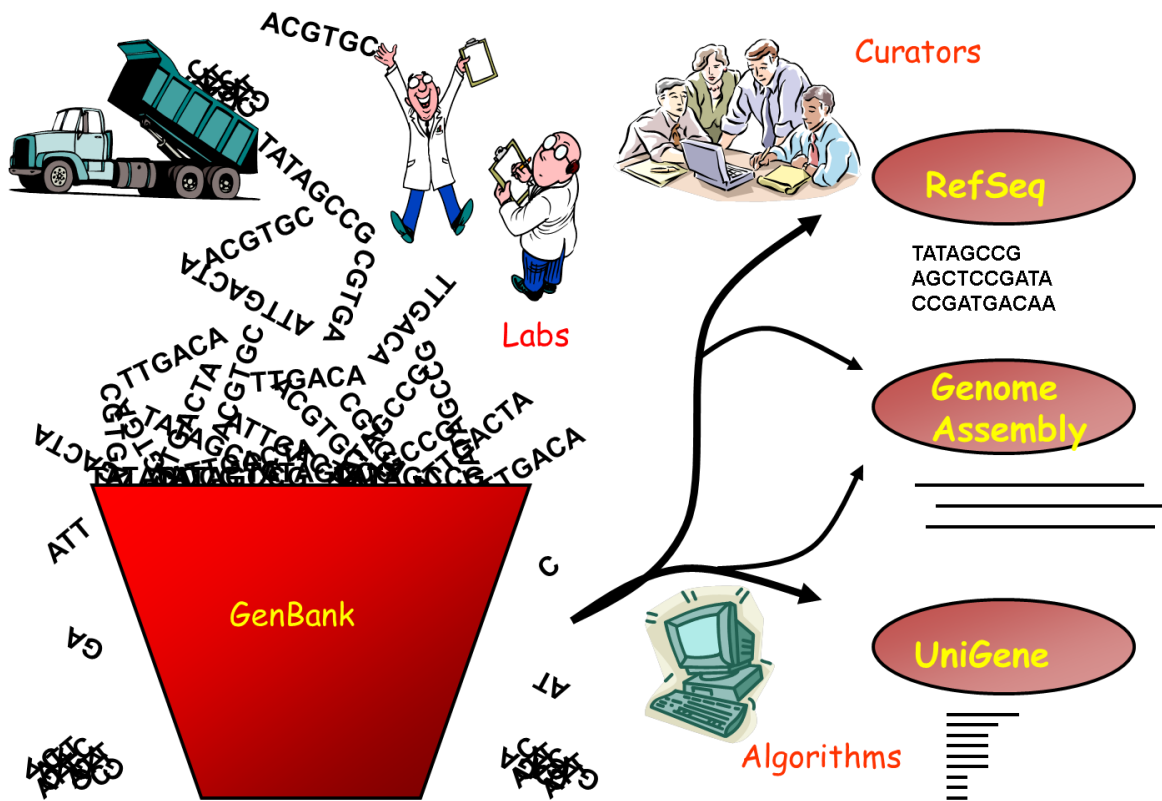
مقدمه

ما در دنیایی زندگی می‌کنیم که در آن حدود یک و نیم 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 و توالی آنها و مانند آن می‌باشد. از پردازش این داده‌ها، 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.

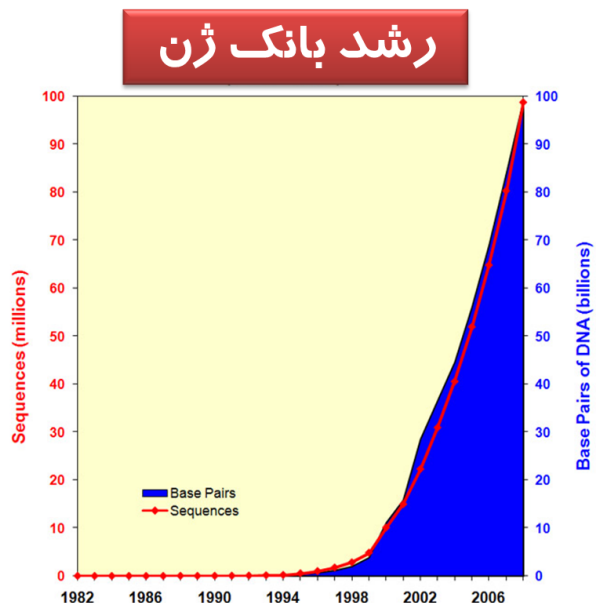


داده پردازی زیستی از اشتراک فعالیت های مرتبط با تولید هر گونه اطلاعات زیستی (از جمله پروژه های ژنوم)، ایجاد پایگاه های اطلاعاتی و ابزارهای مربوط، تولید نرم افزارها و ساخت سخت افزارهای مورد نیاز حاصل می شود. بازیابی اطلاعات حاصل، تجزیه و تحلیل و بالاخره تفسیر آنها نتیجه این فعالیت هاست که تاثیر عمیقی در رشد و پیشرفت علوم زیستی و رشته های مرتبط داشته است.

در حال حاضر، با تلاش های 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 می شود. به طوری که در سال ۱۹۷۷ میلادی تنها توالی یک ژن (گلوبین خرگوش) را می دانستیم. اکنون بیش از ۸۰ میلیون رکورد در این بانک وجود دارد. شمایی از رشد اطلاعات در این بانک در شکل زیر آورده شده است.

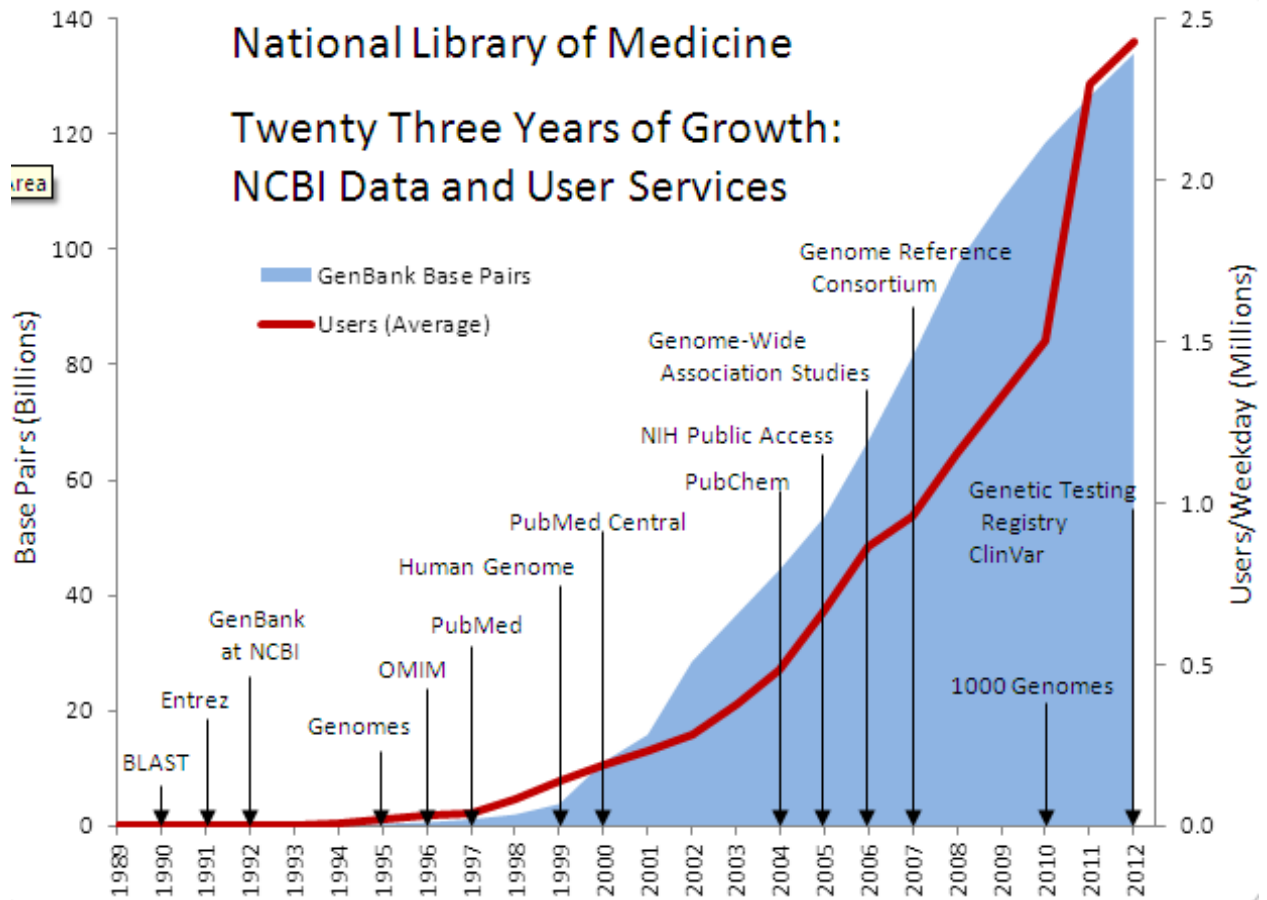


Year	Base Pairs	Sequences
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273
1995	384939485	555694
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. هم‌اکنون 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 می‌رود 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 داده‌پردازی زیستی و پروژه‌های 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 در سایت 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 پایگاه‌های 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 پایگاه‌های اطلاعاتی زیستی وجود دارد. از جمله:

- پایگاه‌های اطلاعاتی اولیه شامل توالی‌های 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 اسیدهای آمینه نوشته می‌شوند (جدول زیر).

پایگاه‌های اطلاعات مربوط به کربوهیدرات‌ها 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

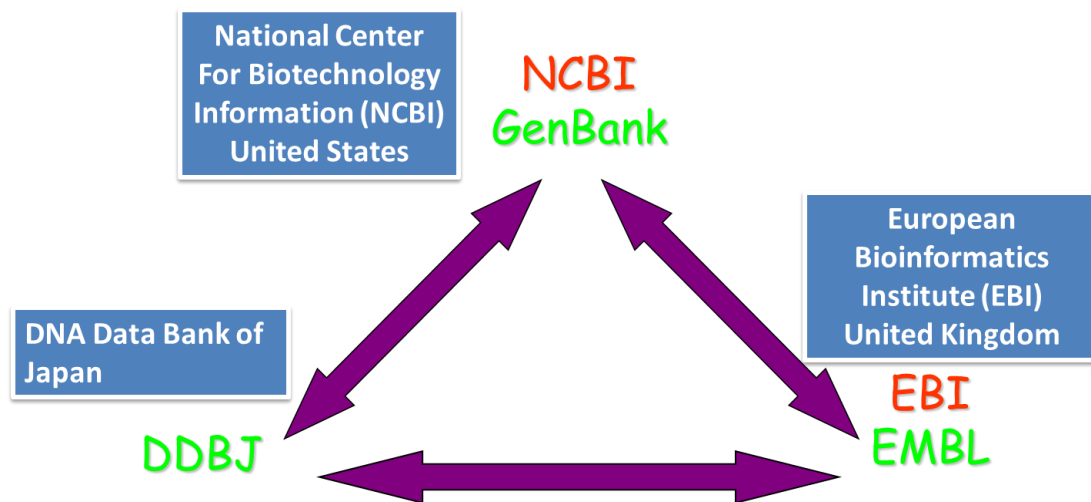
		Second base							
		T		C		A		G	
F i r s t b a s e	T	TTT	Phe (F)	TCT	Ser (S)	TAT	Tyr (Y)	TGT	Cys (C)
		TTC	Phe (F)	TCC	Ser (S)	TAC	Tyr (Y)	TGC	Cys (C)
		TTA	Leu (L)	TCA	Ser (S)	TAA	STOP	TGA	STOP
		TTG	Leu (L)	TCG	Ser (S)	TAG	STOP	TGG	Trp (W)
	C	CTT	Leu (L)	CCT	Pro (P)	CAT	His (H)	CGT	Arg (R)
		CTC	Leu (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 توالی‌های نوکلئوتیدی که در آزمایشگاه‌های مختلف تولید می‌شود را جمع‌آوری 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 وجود 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 می‌شوند:

- 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 سال ۱۹۸۲ 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 ۳ میلیون توالی و ۱۴۰۰ گونه جدید به این بانک اطلاعاتی افزوده می‌گردد به طوری که تقریباً 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 داده‌های تکراری و اضافی (مانند توالی‌های 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 بخش های 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 داده ها) بخشی از اطلاعات خود را در قسمت های اختصاصی زیر ذخیره می کند:

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 می دهد.

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 اضافه یک عدد برای نشان دادن دفعات اصلاح آن رکورد)، منشأ توالی و تاکسونومی موجود منشأ، کلید واژه ها، آدرس مقالات مربوط می باشد. شماره دسترسی مهمترین جزء این قسمت است زیرا تنها این شماره یگانه بوده و بهترین گزینه برای مراجعه مجدد به این رکورد می باشد.

بخش ویژگی‌ها یا Features

در این بخش اطلاعاتی که در مورد ویژگی‌های توالی است می‌آید. طول توالی، محل توالی کدکننده پروتئین (CDS¹) 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

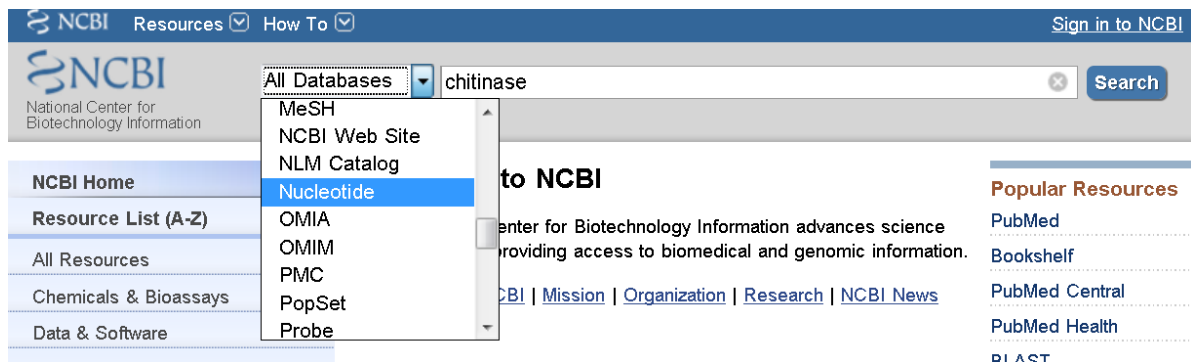
با ارائه آماری از انواع نوکلئوتید و پس از کلمه Origin توالی شروع می‌شود. اعداد در سمت چپ توالی موقعیت باز کنار آن در طول توالی را نشان می‌دهد. توالی ممکن است به کوچکی یک اولیگونوکلئوتید (بیش از ۲۰ جفت باز) در قسمت STS تا به بزرگی میلیون‌ها باز (توالی‌های کروموزم‌های به دست آمده از پروژه‌های ژنوم) باشد.

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 Advance 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

Nucleotide Nucleotide chitinase Save search Limit **Advanced** Help

Display Settings: Summary, 20 per page, Sorted by Default order **Send to:** **Filter your results:**

Found 32589 nucleotide sequences. Nucleotide (28099) EST (4486) GSS (4)

Results: 1 to 20 of 28099 << First < Prev Page 1 of 1405 Next > Last >>

[Allium sativum chitinase mRNA, 3' end](#)
 1. 1,085 bp linear mRNA
 Accession: M94105.1 GI: 166344
[GenBank](#) [FASTA](#) [Graphics](#)

[Allium sativum chitinase mRNA, 3' end](#)
 2. 1,007 bp linear mRNA
 Accession: M94106.1 GI: 166342
[GenBank](#) [FASTA](#) [Graphics](#)

Filter your results:
 All (28099)
[Bacteria \(17667\)](#)
[INSDC \(GenBank\) \(16226\)](#)
[mRNA \(4520\)](#)
[RefSeq \(11841\)](#)
[Manage Filters](#)

Top Organisms [Tree]
 uncultured bacterium (1338)
 Vibrio cholerae (1317)
 Propionibacterium acnes (533)
 uncultured

Nucleotide Advanced Search Builder

(chitinase[Title] AND zebrafish[Organism])

[Edit](#)

Builder

Title [Show index list](#)

AND Organism [Show index list](#)

AND All Fields [Show index list](#)

or [Add to history](#)

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).

Nucleotide Advanced Search Builder

[Edit](#)

Builder

[Show index list](#)

[Show index list](#)

to [Show index list](#)

[Show index list](#)

or [Add to history](#)

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.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide ((chitinase[Title] AND zebrafish[Organism]) AND ("2012"[Modification Date] : "3000"[Modification Date])) Search

Save search Limits Advanced Help

Display Settings: Summary, 20 per page, Sorted by Default order

Send to: **Filter your results:**

All (7)

Bacteria (0)

[INSDC \(GenBank\) \(1\)](#)

[mRNA \(7\)](#)

[RefSeq \(6\)](#)

[Manage Filters](#)

Results: 7

[Danio rerio chitinase, acidic.3 \(chia.3\), mRNA](#)

1. 1,746 bp linear mRNA

Accession: NM_213213.1 GI: 47086016

[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

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.

Display Settings: GenBank

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

NCBI Reference Sequence: NM_213213.1
[FASTA](#) [Graphics](#)

Go to:

LOCUS NM_213213 1746 bp mRNA linear VRT 24-AUG-2013

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

ACCESSION **NM_213213** NM_213498

VERSION NM_213213.1 GI:47086016

KEYWORDS RefSeq.

SOURCE *Danio rerio* (zebrafish)

ORGANISM [Danio rerio](#)
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
 Cypriniformes; Cyprinidae; Danio.

REFERENCE 1 (bases 1 to 1746)
 AUTHORS Hussain,M. and Wilson,J.B.
 TITLE New Paralogues and Revised Time Line in the Expansion of the Vertebrate GH18 Family
 JOURNAL J. Mol. Evol. (2013) In press
 PUBMED [23558346](#)
 REMARK Publication Status: Available-Online prior to print

REFERENCE 2 (bases 1 to 1746)
 AUTHORS Peterson,S.M., Zhang,J., Weber,G. and Freeman,J.L.
 TITLE Global gene expression analysis reveals dynamic and developmental stage-dependent enrichment of lead-induced neurological gene alterations
 JOURNAL Environ. Health Perspect. 119 (5), 615-621 (2011)
 PUBMED [21147602](#)

REFERENCE 3 (bases 1 to 1746)
 AUTHORS Jima,D.D., Shah,R.N., Orcutt,T.M., Joshi,D., Law,J.M., Litman,G.W.,

شماره دسترسی

منبع، نویسندگان، اطلاعات مجله و مقاله مربوط به توالی

زمان تسلیم اطلاعات یا آخرین تغییرات

طول توالی نوع مولکول

VRT (other vertebrate)

FEATURES

source Location/Qualifiers
 1..1746
 /organism="Danio rerio"
 /mol_type="mRNA"
 /db_xref="taxon:7955"
 /chromosome="11"
 /map="11"

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; zgc:77912"
 /codon_start=1
 /product="chitinase, acidic.3 precursor"
 /protein_id="NP_998378.1"
 /db_xref="GI:47086017"
 /db_xref="GeneID:406819"
 /db_xref="ZFIN:ZDB-GENE-040426-2891"

translation="WGRLTLIAGLSLVLCHVAFSMEMACYFTNWSQYRPGIGKYTPAVDPYLCTHLIYAFSIIQNRNELVTYEWNDLTKAFNELKKNPTLTKLLAVGGWNFCSAQFSINVSNPANRKTFIQSTIKFLRTHGFDGLDLDMEYPGARGSPPEKQRFLLCKELVAAYEAESKATGNPQLNMLTAAVSAAGKGTIDGGEYIAEIAKYLNFIMVHTYDFHGTWERFTGHNSPLYQGSKDEGLIYFNTDYANRYWRDNGTPVEKLRMGFAAYGRTFRLTSDTSVSGAPASGPASAGTYTREGFWSYEICGFLEGTIIQWIDQKVPYATKNSEWVGFDTKESYETKVRYLKDKNFGGAFVWALDLDDFAGQFCQGNHPLMAHLRNLDDIELPPMPSSTTPKPGQSTTRPTTTTTTTTHAPGPGFCNGKPDGLYAHNPDPNKYYSCAGGHTFVEKCAVGTVFDDSKCCVWPKP"

sig_peptide
 28..87
 /gene="chia.3"

توالی کدکننده Coding Sequence

ترجمه توالی نوکلئوتیدی به اسید آمینه

شماره دسترسی به توالی پروتئین مربوطه

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;

```

کدون آغاز

28

CDS

28..1443

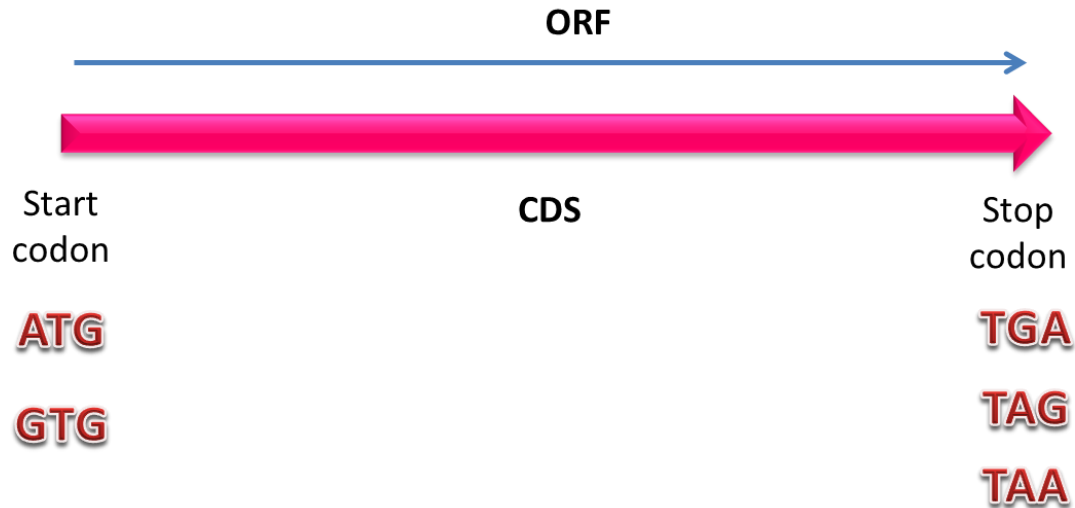
1443

کدون پایان

```

1 gaggtacccc agaaaattgg cattaagaatg gggagactta cacttatagc aggtttgagc
61 ttggtgctct gccatgtcgc cttttccatg gaaatggcct gctacttcac caactggctc
121 caatatagac ctggaattgg aaaatatact cctgcaaagc tcgaccctta cctttgcaca
181 caccttatct atgctttttc aatcatcaac caaaggaatg agcttgcac atatgagtgg
241 aacgatgaaa ctctatacaa ggccttcaat gaactcaaga acaaaaaatcc cactcttaag
301 acccttttgg ctgttgagg atggaatttt ggttcggcac agttctccat catggtgtcc
361 aatcctgcaa accgtaaaac atttattcag tctaccatca aattcctgag aactcatgga
421 tttgatggac tggatctgga ctgggaatat cccggagcaa gagggaagtcc acctgaagac
481 aaacaaagat tcactctgct gtgcaaggaa cttgttcag cctatgaggc tgagagtaaa
541 gccactggca atcctcagct tatgtctgacc gctgctgtat cagctggcaa aggcactatt
601 gatgatggat atgagattgc agagattgcc aagtacttga acttcatcaa cgtcatgact
661 tacgacttcc atggcacttg ggaagcattc acaggacaca acagccctct gtaccaaggc
721 tcaaaggatg agggagacct gatctacttc aacaccgact atgctatgag ctactggagg
781 gacaatgaaa cccctgtgga gaaactcaga atgggctttg cagcatacgg tcgcactttc
cagcgttggc gctccagcta gtggacctgc ttcagctgga
961 acaacaattc agtggattga tgaccagaag gtgccctatg ccacaaagaa cagcagtgga
1021 gttggatttg acaccaagga gaggatgaa acgaaggtcc gttatctgaa agacaagaat
1081 tttggtggag ctttcgtttg ggcacttgat ctggatgact ttgctggaca gttctgtagt
1141 caggggaacc atcctctcat ggccatctt cgcaatctt tggatattga attgcctcca
1201 atgecttcaa ctaccactcc taaacctggc caaagcacc caaggccgac cacaaccaca
1261 actaccacca ctcatgctcc aggaccagga ttctgtaatg ggaagccaga tggactctat
1321 gctcacccta atgaccccaa caaatattac agctgtgctg gaggtcatac cttcgtggaa
1381 aaatgtgctg taggcaccgt gtttgatgac agctgcaagt gctgtgtttg gcccaaacct
1441 tagtcatcat gactcaagaa acttcagaaa aacatttgca aatagtgaaa caagacactg
1501 caaatttctt taagccaaca aatgacagca aaccgtttca tatataagaa cattagtgtg
1561 acctacattt atttactaaa tttgtattat gtctttatta catcttgggtg acgtttatac
1621 tgtaagtgca agtacctctt tttaaatgac tgacaaaaag tttgtttctt gttcgtttct
1681 cttgcttgtg gaaattaata aactgacatc atggagaggg gaaaaaaaaa aaaaaaaaaa
1741 aaaaaa

```



قالب‌های (Formats) نمایش توالی

پس از جستجو، می‌توانید رکورد یا توالی مورد نظر را به اشکال مختلف نمایش دهید. این موضوع هنگامی اهمیت می‌یابد که می‌خواهید تجزیه و تحلیل بیشتر روی توالی به دست آمده با استفاده از نرم‌افزارهای دیگر انجام دهید. اکثر نرم‌افزارها قالب FASTA را می‌پذیرند. در این قالب، سطر اول با علامت > آغاز شده و یک متن حاوی نام توالی به دنبال آن می‌آید. سطرهای بعدی فقط توالی هستند. از گوشه سمت چپ بالای صفحه نمایش توالی می‌توان به فرمت FASTA مربوط به توالی جستجو شده دست پیدا کرد (شکل ۹). و مطابق شکل ۱۰ می‌توان یک توالی را با فرمت FASTA ذخیره کرد.

[Display Settings:](#) GenBank

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

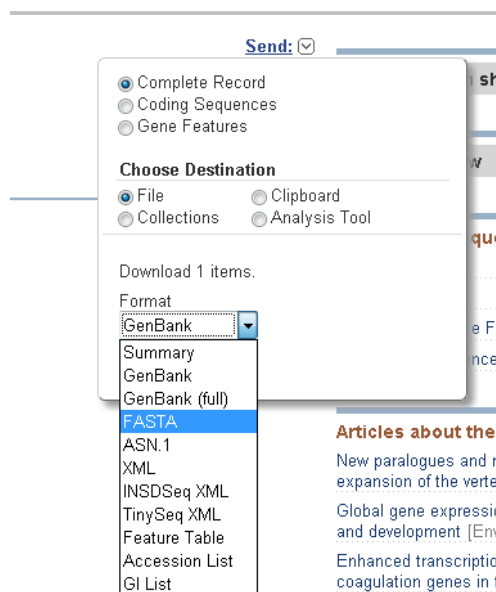
NCBI Reference Sequence: NM_213213.1

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS	NM_213213	1746 bp	mRNA	linear	VRT 24-AUG-2013
DEFINITION	Danio rerio chitinase, acidic.3 (chia.3), mRNA.				
ACCESSION	NM_213213 NM_213498				
VERSION	NM_213213.1 GI:47086016				
KEYWORDS	RefSeq.				

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 رکوردهای GenBank

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

قسمت برجسب‌های 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 می‌آیند.

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

در این قسمت توالی‌های بررسی ژنومی (Genome Survey Sequences) ذخیره می‌شود. توالی‌های GSS کوتاه بوده و یک‌بار توالی‌یابی شده‌اند. این توالی‌ها تصادفی بوده و معمولاً از انتهای کلون‌های کاسمیدی و BAC به دست 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 برجسب 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.gov/SNP>


این پایگاه پلی مورفیسم‌های تک نوکلئوتیدی را ذخیره می‌کند. حذف‌ها و تداخل‌های هر توالی، تکرارهای پلی مورفیک و تنوع میکروساتلیتی در این پایگاه قرار می‌گیرد.

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

در ماه می سال ۲۰۰۰ سه پایگاه اطلاعاتی EMBL, DDBJ 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 بخش‌های 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 می‌شود.

پایگاه داده Refseq

مجموعه توالی‌های مرجع استخراج شده از GenBank را گویند که تصحیح شده و غیر تکراری هستند. یعنی برخلاف GenBank، در پایگاه Refseq 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.



NM_123456	mRNA	
NP_123456	Protein	
NR_123456	RNA	Non-coding transcripts
NG_123456	Genomic	Incomplete genomic region
NT_123456	Genomic	BAC sequence assemblies
NW_123456	Genomic	WGS sequence assemblies
NC_123456	Genomic	Complete genomic molecules
XM_123456	mRNA	Genome Annotation
XR_123456	RNA	Genome Annotation
XP_123456	Protein	Genome Annotation

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 آگزون‌ها و ایترون‌ها را نشان می‌دهد. شکل ۱۱ نمونه‌ای از این نوع جستجو را نشان می‌دهد.

The image shows a screenshot of the NCBI Gene database search results for 'chitinase' in zebrafish. The search dropdown menu is open, showing various database options with 'Gene' selected. Below the search results, a genomic map of Chromosome 11 (NC_007122.5) is shown, highlighting the location of the chitinase gene (chi alpha 3 and chi alpha 2) relative to other genes like zgc:73328, LOC101886411, sox12, and pnp4a.

محدوده‌ای که ژن روی کروموزوم قرار دارد

NC_007122.5 (25040795..25043718)

Chromosome 11 NC_007122.5

[25012993] [25231258]

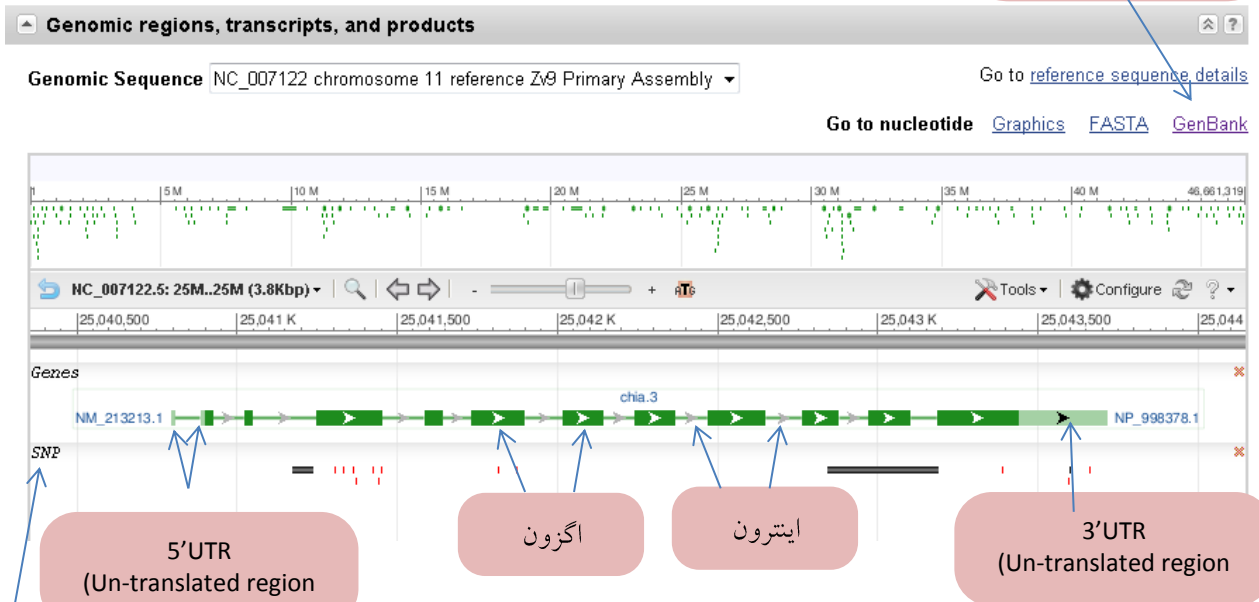
zgc:73328 ← LOC101886411 ← chi α+3 → chi α+2 →

sox12 ← GCG6U → pnp4α →

HOW010G068

شکل ۱۱. جستجو در پایگاه Gene.

برای دستیابی به توالی
ژنومی روی GenBank
کلیک کنید.



بانک مربوط به
جهش‌های
نقطه‌ای
SNP

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 نوکلئوتیدی 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 ناحیه بدون ترجمه سمت 5' گفته می‌شود که می‌توان جایگاه اتصال ریبوزوم (RBS) را به همراه داشته باشد. بر همین اساس محدوده 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).

Change region shown

Whole sequence

Selected region

from: 100 to: 5000

Update View

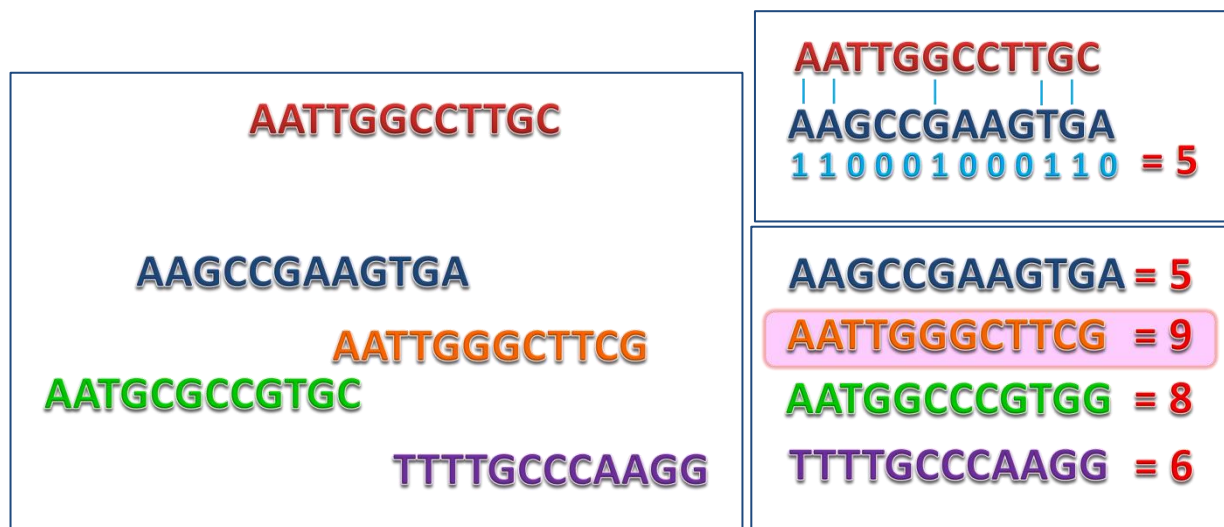
محدوده نمایش توالی‌های ژنومی را می‌توان بر حسب نیاز مطابق شکل بالا تغییر داد تا به توالی‌های اطراف توالی مورد جستجو دست پیدا کرد.

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 می‌آوردند. 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 می‌شود. به عنوان مثال، ساده‌ترین نوع امتیاز دهی این گونه می‌تواند باشد که اگر دو بازی که زیر هم قرار می‌گیرند 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. تفاوت راه حل‌ها و الگوریتم‌ها با هم در این است که جواب کدام یک به واقعیت 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 سؤالاتی 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 سؤالات 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 رابطه 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 دات پلات 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

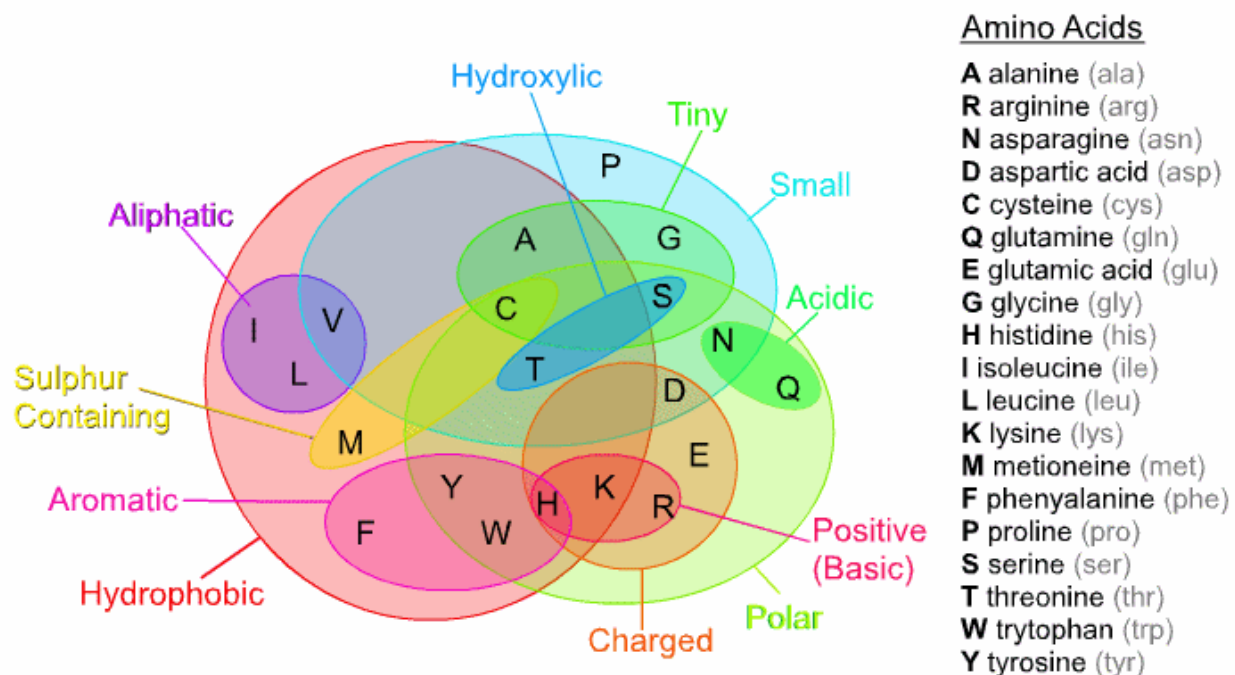
مقایسه همردیفی

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 رشته‌ای DNA 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، همه 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.

نکته دیگر این است که در هم‌ردیفی‌های بالا فرض، برابر بودن طول توالی‌هاست. در حالی که در تکامل توالی‌ها هم پدیده‌ی اضافه شدن را داریم و هم پدیده حذف اتفاق می‌افتد. بنابراین ممکن است در بسیاری از موارد بخواهیم دو توالی را با هم مقایسه کنیم که دارای طول یکسانی نباشند.



جمع‌بندی این مقدمات نشان می‌دهد می‌توان با کمی نمودن (امتیازدهی) نتایج هم‌ردیفی، آنها را باهم مقایسه نمود. البته برای کمی نمودن 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 و یک کفایت نمی‌کند و جایگزینی نوکلئوتیدها یا اسید آمینه‌ها با یکدیگر امتیاز منفی یا مثبت مساوی ندارند، متخصصین امر در پی تهیه جداول امتیازدهی جایگزینی‌ها (Substitution Scoring Matrices) بوده‌اند. به طوری که تا حد 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 متخصصین 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 +5 و برای غیر جفت شدگی امتیاز -4 در نظر گرفته شده است. سایر حروف در صورت وجود انتخاب برای دو نوکلئوتید یا بیشتر در هر موقعیت از توالی کاربرد دارند.

Matrix Structure: Nucleotides

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-1	-1	-4	-1	-2
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-1	-1	-1	-4	-2
S	-4	-4	1	1	5	-4	-2	-2	-2	-2	-1	-1	-3	-3	-1
W	1	1	-4	-4	-4	5	-2	-2	-2	-2	-3	-3	-1	-1	-1
R	1	-4	1	-4	-2	-2	5	-4	-2	-2	-3	-1	-3	-1	-1
Y	-4	1	-4	1	-2	-2	-4	5	-2	-2	-1	-5	-1	-3	-1
K	-4	1	1	-4	-2	-2	-2	-2	5	-4	-1	-3	-3	-1	-1
M	1	-4	-4	1	-2	-2	-2	-2	-4	5	-3	-1	-1	-3	-1
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	5	-4	-2	-2	-1
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-4	5	-2	-2	-1
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-2	-2	5	-1	-1
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-2	-2	-1	5	-1
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	5

- Simple match/mismatch scoring scheme:

Match	+ 5
Mismatch	- 4
- Assumes each nucleotide occurs 25% of the time

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 پروتئینی انجام و گروه‌بندی شد. سپس فراوانی جایگزینی یک اسید آمینه با اسید آمینه دیگر در فرمول زیر به کار گرفته شد:

$$PAM_n = \log \text{Probability of one substitution} / \text{Probability of occurring by chance} * 100$$

در این فرمول یک واحد PAM (Point Accepted Mutation) معادل یک تغییر در یک توالی صدماتی 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 آن 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	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	0	-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	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	0	-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	0	-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	0	-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	0	-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	0	-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	0	-8
I	-1	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	0	0	-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	0	-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	0	-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	0	-8
F	-3	-4	-3	-6	-4	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	0	0	-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	0	-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	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	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	0	-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	0	-8
V	0	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	0	0	-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	0	-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	0	-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	0	-8
*	-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

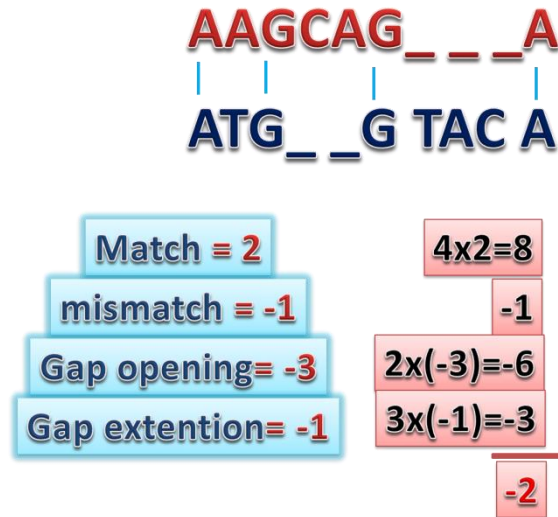
بعدها جداول دیگری تدوین شدند که از یک نوع اصول پیروی می‌کردند. با این تفاوت که فراوانی جایگزینی‌ها تنها در مناطق حفاظت شده (Conserved Blocks) برای ساختن جدول، محاسبه می‌شدند. در آن هنگام، ۲۰۰۰ بلوک از ۵۰۰ خانواده پروتئینی در نظر گرفته شد. به طور مشابهی، فرمول زیر بکار رفت:

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

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

امتیازدهی حذف و اضافه نمودن توالی ها

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



یک مثال از سیستم امتیازدهی فرضی.

این که فواصل را در هم ردیفی چگونه محاسبه کنیم یکی از مبهم ترین مساله ها در هم ردیفی توالی ها است. به طور معمول، جریمه هایی را که برای فواصل در نظر می گیرند به صورت محلی اعمال می شود. یعنی جریمه استفاده از هر فاصله مستقل از فواصل دیگری است که ممکن است در جاهای دیگری از هم ردیفی اتفاق بیافتد. در همه برنامه ها، برای فواصل دو نوع امتیاز منفی در نظر گرفته می شود.

۱. جریمه باز کردن توالی: (GAP)

در طبیعت، هر اضافه نمودن توالی مستلزم صرف انرژی بوده و مورد انتخاب طبیعی قرار خواهد گرفت. بنابراین در الگوریتم ها برای هم ردیفی بهینه توالی ها امتیاز منفی نسبتاً بزرگی (مثلاً -۱۱) برای ایجاد فاصله در نظر گرفته می شود (در مثال بالا این امتیاز منفی ۳ در نظر گرفته شده بود).

¹Gap opening penalty

۲. جریمه بسط یک فاصله (GEP)¹

از این جریمه برای ورود نوکلئوتید یا اسید آمینه در محلی که قبلاً فاصله ایجاد شده است، استفاده می‌شود. میزان این جریمه از GOP کمتر فرض می‌شود ولی در تعداد آنها ضرب می‌شود (مثلاً 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 گانه 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 الگوریتم‌های نوشته شده برای هم‌ردیفی دوگانه صورت می‌گیرد.

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

¹ Gap extension penalty

BLAST چگونه کار می کند

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 (high-scoring segment pairs) HSPs نامیده می شوند. برای این کار برنامه BLAST از روش Dynamic Programming استفاده می کند.

در روش dynamic programming برای حل یک مشکل بزرگ، آن را به چند مشکل کوچک تجزیه می کنند. پس از یافتن پاسخ مناسب مشکلات کوچک، آنها را کنار هم چیده و راهی برای پاسخ به مشکل بزرگ پیدا می کنند. 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 اندازه کلمات برای توالی‌های آمینو اسیدی ۳ و برای توالی‌های نوکلئوتیدی ۱۱ تنظیم شده است. سپس از بین این کلمات انتهایی انتخاب می شوند 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.

توالی مورد نظر
(Query)

MNPLSSSGQPHTLM

MNP

NPL

PLS

LSS

SSS

SGQ

GQP

QPH

PHT

HTL

TLM

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

توالی‌های مشابه ممکن
(subjects)

MNGPLSSSGQTSTSPH
LSS
PLSSSSGQ

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

PLSSSGQ
| | | | | | | |
PLSSSGQ

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 هر یک 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. نتیجه جستجو، توالی‌های پروتئینی 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، توالی مورد تقاضا نوکلئوتیدی است که در ۶ قالب خواندنی (ORF) ترجمه شده و به صورت توالی پروتئینی در پایگاه توالی‌های پروتئینی جستجو می‌شود. نتیجه 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 می‌شود 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 کننده آن را شناسایی کنیم.

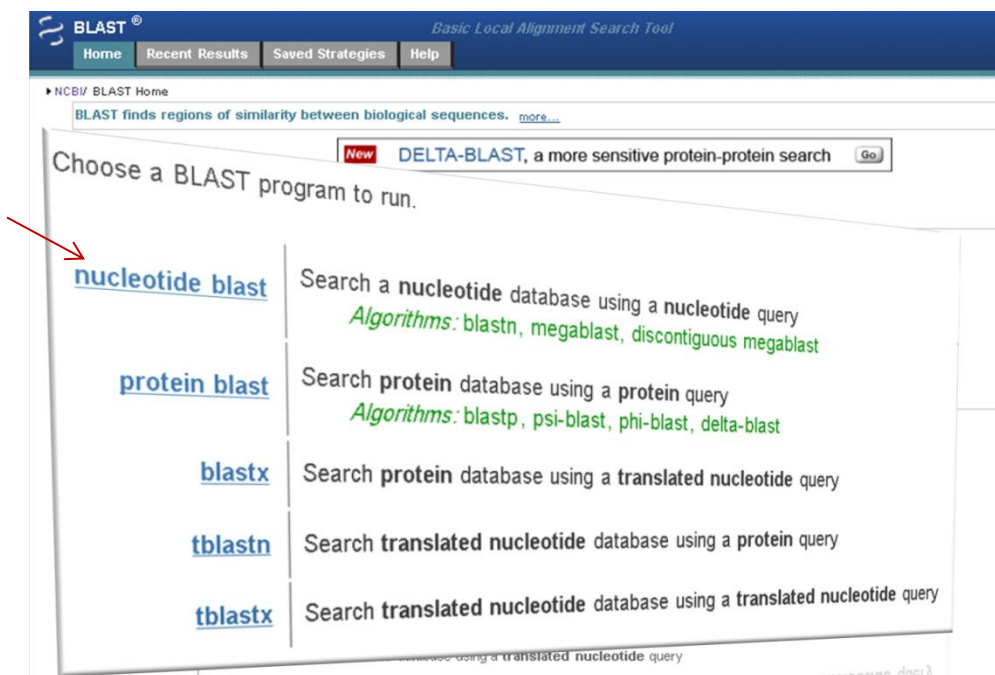
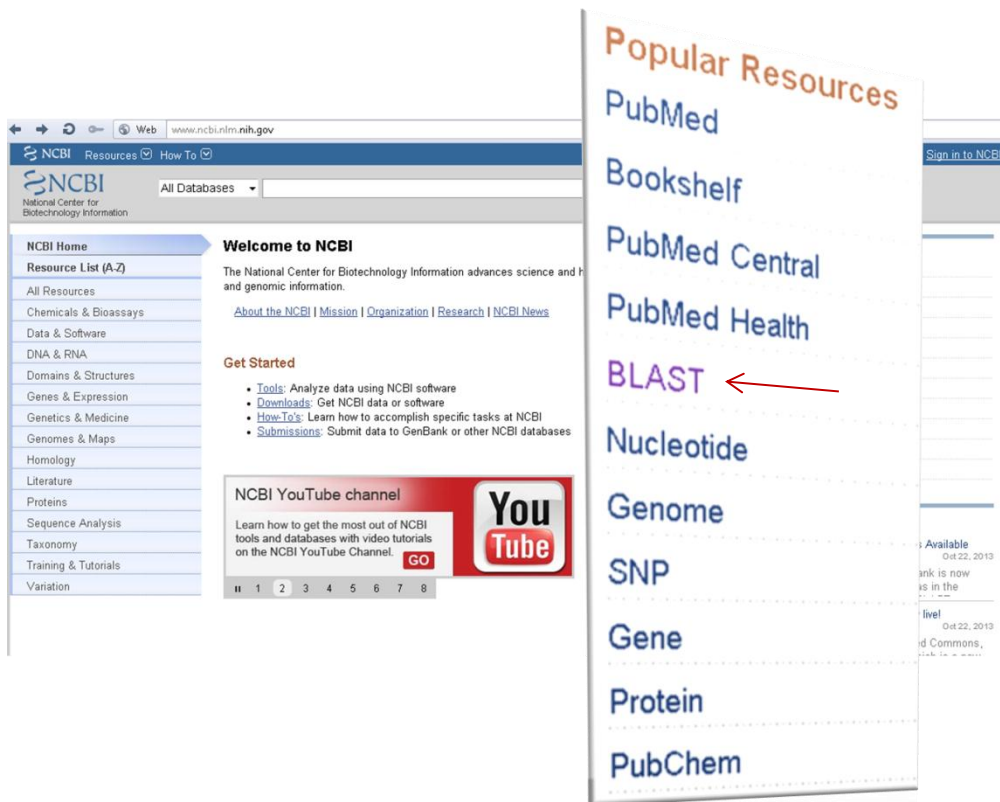
- **tBLASTX**

در این نوع BLAST، توالی مورد تقاضا، نوکلئوتیدی است که در ۶ قالب خواندنی به پروتئین ترجمه می‌شود و در پایگاه توالی‌های نوکلئوتیدی که آن نیز در ۶ قالب خواندنی به پروتئین ترجمه می‌شود مورد جستجو قرار می‌گیرد. این نوع جستجو به ویژه در مطالعات EST به کار می‌رود.

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

```
GAGGTACCCAGAAAAATGGCATTAGAATGGGGAGACTTACACTTATAGCAGGTTTGAGCTGGTGCTCT
GCCATGTGCCTTTTCCATGGAAATGGCCGTACTTCCACCACTGGTCCCAATATAGACC TGGAAATGG
AAAAATATACTCCTGCAAAATGTCGACCCTTACCTTTGCACACACCTTATCTATGCTTTTCAATCATCAAC
CAAAGGAATGAGCTTGTACATATG AAGCCTTCAATGAACTCAAGA
ACAAAAATCCCACTCTTAAGACCCT TGGTTCGGCACAGTTCTCCAT
CATGGTGTCCAATCCTGCAAACCGT AAATCCTGAGAACTCATGGA
TTTGATGGACTGGATCTGGACTGGG CACCTGAAGACAAACAAGAT
TCACTCTGCTGTGCAAGGAACCTTGT AGCCACTGGCAATCCTCAGCT
TATGCTGACCGCTGTGTATCAGCT TATGAGATTGCAGAGATGGCC
AAGTACTTGAACCTCATCAACGTCA GGGAGCGATTACAGGACACA
ACAGCCCTCTGTACCAAGGCTCAAA CAACACCGACTATGCTATGCG
CTACTGGAGGGACAAATGGAACCCCT GCAGCATACGGTCGCACTTCT
CGTCTGACATCTTCAATACCAAGCG CTTCAGCTGGAACTACACTC
GCGAGGCTGGATTCTGGTCTTACTA AACAACAATTCAGTGGATTGA
TGACCAGAAGTGCCCTATGCCACA GACACCAAGGAGAGTTATGAA
ACGAAGTCCCGTTATCTGAAAGACA GGGCACTTGATCTGGATGACT
TTGCTGGACAGTTCTGTAGTCAGGG TCGCAATCTTCTGGATATTGA
ATTGCCCTCAATGCCTTCAACTACC ACAAGGCGACCACAACCACA
ACTACCACCACTCATGCTCCAGGACCAGGATTCTGTAATGGGAAGCCAGATGGACTCTATGCTCACCCTA
ATGACCCCAACAATAT TACAGCTGTGCTGGAGGTCATACCTTCGTGGAAAAATGTGCTGTAGGCACCGT
GTTTGATGACAGCTGCAAGTGTGTGTTGGCCCAAACCTTAGTGATCATGACTCAAGAAACTTCAGAAA
AACATTTGCAAAATAGTGAACACAGACACTGCAAAATTTCTTAAAGCCAACAATGACAGCAAAACCGTTTCA
TATATAAGAACATTAGTGTAACTACATTTATTTACTAAATTTGATATATGCTTTATACATCTTGGTG
ACGTTTATACTGTAAAGTCAAGTACCTCTTTTAAATGACTGACAAAAAGTTGTTTCTGTTGTTCTTCT
CCTTGCTTGTGAAATTAATAAATGACATCATGGAGAGGGGAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

- Search
- Search With
- Copy
- Copy to Note
- Speak
- Dictionary
- Encyclopedia
- Translate
- Go to Web Address
- Send by Mail

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
CACATAGCTCAGTTCCCATAAAAGGGCTGGTTTGCCGCGTGGGGAGTGGAGTGGGACAGGTATATAAG
GAAGTACAGGGCTGGGGAGAGGCCCCGTCTAGGTAGCTGGCACCAGGAGCCGTGGGCAAGGGAAGAGG
CCACACCCCTGCCCTGCTCTGCTGCAGCCAGAAATGGGTGTGAAGGCGTCTCAAAACAGGCTTTGTGGTCTG
GTGCTGCTCCAGTGTCTCTGCTACAAAATGGTCTGCTACTACACAGCTGGTCCAGTACCGGGAAAG
GCGATGGGAGCTGCTCCAGATGGCCCTGACCGCTTCTCTGTACCCACATCATCTACAGCTTTGCCAA
TATAAGCAACGATCACATCGACACCTGGGAGTGGAAATGATGTGACGCTC
AAGAACAAGAACCCCAACCTGAAGACTCTCTTGTCTGTGCGGAGGATGGAA
CCAAGATAGCTCCAAACCCAGAGTCCGCGACTTTCATCAAGTCAAGT
TGGCTTTGATGGGCTGGACCTTGCCTGGCTCTACCCCTGGACGGAGAGAC
ATCAAGGAAATGAAGCCGAATTTATAAAGGAAGCCAGCCAGGGAAAAA
CACTGCTGCGGGGAAGGTCACCATTGACAGCAGCTATGACATTGCCAA
CATTAGCATCATGACCTACGATTTTCATGGAGCTGGCGTGGGACACAG
CGAGGTCAGGAGGATGCAAGTCTGACAGATTTCAGCAACACTGACTATG
TGGGGCTCCTGCCAGTAAAGTGGTGGTGGGATCCCCACCTTCGGGAG
TGAGACTGGTGTGGAGCCCAATCTCAGGACCGGGAATTCAGAGCCGG
CTTGCCACTATGAGATCTGTGACTTCTCCGCGGAGCCACAGTCCATAG
CCTATGCCACCAAGGGCAACAGTGGGTAGGATACGACGACCAGGAAAG
CCTGAAGGACAGGCAAGTGGCGGGCGCCATGGTATGGGCTTGGACCTG
TGGGGCAGGATCTGCGCTTCCCTCTCACC AATGCCATCAAGGATGAC
TTCTGCACACAGCACGGGGCCAAAGATGCCCGCTCCCCCTCTGGCTCC
CCTGCCCTGCTGAGTCCAGGCTGAGCCTCAAGTCTCCCTCCCTTGGGGCTATGCGAGGTTCCACAAAC
ACAGATTTGAGCTCAGCCCTGGTGGGCAGAGAGGTAGGGATGGGGCTGTGGGGATAGTGAAGCATCGCAA
TGTAAAGACTCGGGATTAGTACACACTTGTGATTAATGGAAATGTTTACAGATCCCCAAGCTTGGCAAGG
GAATTTCTCAACTCCTTGCCCCCAGCCCTCCTTATCAAAGGACACATTTTGGCAAGCTCTATCACC
AGGAGCCAAACATCCTCAAGACACAGTGACCATACTAATATACCCCTGCAAGGCCAGCTTGAAGC
TTCACTTAGGAACGTAATCGTGTCCCCTATCCTACTTCCCTTCTAATCCACAGCTGCTCAATAAAGT
ACAAGAGCTTAACAGTGAATAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

- Search
- Search With
- Copy
- Copy to Note
- Speak
- Dictionary
- Encyclopedia
- Translate
- Go to Web Address
- Send by Mail

blastn blastp blastx blastz

BLASTN program c

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

GTTTGATGACAGCTGCAAGTGCATGTTTGGCCCAACCTTAGTGCATGACTCAAGAACTTCAGAAA
AACATTTGC AATAGTGAAC AAGACACTGC AATTTCC TTAGCC AAC AATGACAGC AACC GTTC A
TATATAAGAACATTAGTGAACCTACATTATTACTAAATTTGTATTATGCTTTATTACATCTGGTG
ACGTTTATACGTAAGTGC AAGTACCCTTTTTAAATGACTGACAAAAAGTTTGTTCCTTGGTTCTCT
CCTTGC TTGTAARTTAATAACTGACATCATGGAGAGGGGAAAAAAAAAAAAAAAAAAAAAAAA

From

To

Or, upload file

File input field

Choose...

Job Title

gi|47086016|ref|NM_213213.1| Dab1b rerb ckbbase...

Enter a description for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Subject subrange

TGTAAGACTCGGGATTAGTACACACTTGTGATTAATGGAATGTTTACAGATCCCCAAGCCTGGCAAGG
GAATTTCTTC AAC TCCCTGCCCCCAGCCCTCCTATC AAGGAC ACC ATTTGGC AAGCTCTATCACCA
AGGAGCC AAC ATCCTAC AAGACAC AGTGACCATACTAATTATACCCCTGC AAGCCCAGCTTGAACC
TTC ACTTAGG AAGTARTCGTGTCCCTATCC TACTCCCTTCCTAATTCCACAGCTGCTCATTAAGT
ACAAGAGCTTAACAGTGA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

From

To

Or, upload file

File input field

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)

Show results in a new window

+ Algorithm parameters

Note: Parameter values that differ from the default are highlighted

NCBI/BLAST/blastn suite-2sequences/Formatting Results - 6YPB1S9J11N

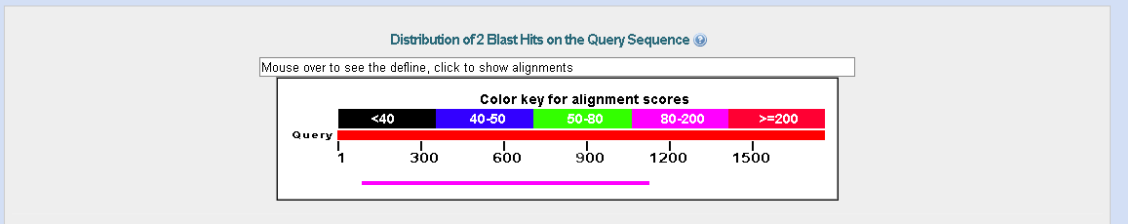
[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#) [YouTube](#) [How to read this page](#) [Blast report description](#)

gi|47086016|ref|NM_213213.1| Danio rerio chitinase,... Blast 2 sequences

RID	6YPB1S9J11N (Expires on 10-30 14:44 pm)	Subject ID	Id 14945
Query ID	Id 14943	Description	gi 144226250 ref NM_001276.2 Homo sapiens chitinase 3-like 1 (cartilage glycoprotein-39) (CHI3L1), mRNA
Description	gi 47086016 ref NM_213213.1 Danio rerio chitinase, acidic.3 (chia.3), mRNA	Molecule type	nucleic acid
Molecule type	nucleic acid	Subject Length	1867
Query Length	1746	Program	BLASTN 2.2.28+ Citation

Other reports: [Search Summary](#) [Taxonomy reports](#)

Graphic Summary



Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

		Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	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



Max score	Total score	Query cover	E value	Ident	Accession
91.5	114	58%	9e-22	64%	30993

gij144228250[ref|NM_001276.2| Homo sapiens chitinase 3-like 1 (cartilage glycoprotein-39) (CHI3L1), mRNA
 Sequence ID: lc|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---CCGGGAAGCGATGGGAG-CTG			293
Query 152	CTGC--AAATGTC---GACCCTTACCTTTGCACACACCTTATCTAT-GCTTTTTCAATCA			205
Sbjct 294	CTTCCCAGATGCCCTTGACCGCTCCTCTGTACCCACATCATCTACAGCTTTGCCAAT-A			352
Query 206	TCAACCAAAGGAATGAGCTTGTACATATGAGTGGAAACGATGAAACTCTATACAAGGCCT			265
Sbjct 353	T-AAGCAACG--ATCACATCGACACCTGGGAGTGGAAATGATGTGACGCTCTAC--GGCAT			407
Query 266	--TCAATGAACCTCAAGAACAAAAATCCCACCTCTTAAGACCCCTTTGGCTGTGGAGGATG			323
Sbjct 408	GCTCAACACACTCAAGAACAGGAACCCCAACCTGAAGACTCTCTGTCTGTGGAGGATG			467
Query 324	GAATTTGGTTCCGGCAGAG-TTCTCCATCATGGTGTCCAACTCTGCAAAACCGTAAAACAT			382
Sbjct 468	GAACCTTTGGTCT-CAAAAGATTTCCAAAGATAGCCTCCAAACCCAGAGTCCGCCGACTT			526
Query 383	TTATTCAGTC--TACCATCAAATTCCTGAGAACTCATGGATTGATGGACTGGATCTGGA			440
Sbjct 527	TATCAAGTCACTACCGCCA--TTCTGCGCACCCATGGCTTGATGGGCTGGACTTGTG			584
Query 441	CTGGGAATATCCCGGAGCAAGAGGAAGTCCACCTGAAGACAAACAAGATTCACTCTGCT			500
Sbjct 585	CTGGCTCTACCTTGGACGGAGAG-----ACAAACAGCATTTTACCACCT			629
Query 501	GTGCAAGGAACCTTGTGTCAGCCT--ATGAGGCTGAGAGTAAGCCACTGGCAATCCTCAG			558
Sbjct 630	AATCAAGGAAATGAAGGCCGAATTTATAAAG--GA-AGCCAGCCA--GGGAAAAAGCAG			684
Query 559	CTTATGCTGACCGCTGCTATCAGCTGGCAAAGGCACATTTGATGATGGATATGAGATT			618
Sbjct 685	CTCTGCTCAGCGCAGCACTGTCTGCGGGAAAGGTACCATTGACAGCAGCTATGACATT			744
Query 619	GCAGAGATTGCCAAGTACTTGAACCTCATCAACGTCATGACTTACGACTTCCATGGCACT			678
Sbjct 745	GCCAAAGTATCCCAACACCTGGATTTCATTAGCATCATGACCTACGATTTTCATGGAGCC			804
Query 679	TGG-GAGCGATTACAGGACACAACAGCCCTCTGTACCAAGGCTCAA-AGGATGAGGGAG			736
Sbjct 805	TGGCGTGGGAC-CACAGGCCATCACAGTCCCTGTTCGAGG-TCAGGAGGATGCA--AG			860
Query 737	ACCTGATCTACT--CAACACCGACTATGCTATGCTGCTAC-TGGAGGGACAATGGAACCC			793
Sbjct 861	TCTGACAGATTCAAGCAACACTGACTATGCTGTGGGGTACATGTTGAGGCTGGGGG-CTC			919
Query 794	CTGTGGAGAAACTCAGAAATGGGCTTGCAGCATACGGTCGCACTTTCCGTCGACATCTT			853
Sbjct 920	CTGCCAGTAAGCTGGTGGATGGGCATCCCCACCTTCGGGAGGAGCTTCACTCTGGCTTCTT			979
Query 854	CAGATACCAGCGTTGGAGCTCCAGCTAGTGGACCTGC--TTCAAGCTGGAACCTACACTCG			911
Sbjct 980	CTGAGACTGGTGTGGAGCCCAATCTCAGGACCGGAATTC--CAGGCCGGTTCACCAA			1037
Query 912	CGAGGCTGG-ATTCTGGTCTTACTATGAGATCTGTGGATTCTGGAGGGAACAACAATTC			970
Sbjct 1038	GGAGGCAAGGACCTTGCCCT-ACTATGAGATCTGTGACTTCTCCGCGGAGCCACAGTCC			1096
Query 971	AGTGGATTGATGACCAGAAGGTGCCCTATGCCACAAGAAGCAGCGAGTGGGTGGATTGG			1030
Sbjct 1097	ATAGAATCCTCGGCCAGCAGGTCCTTATGCCACCAAGGGCAACAGTGGGTAGGATACG			1156
Query 1031	ACACCAAGGAGAGTTATGAAACGAAGGTCGGTTATCTGAAAGACAAGAAATTTGGTGGAG			1090
Sbjct 1157	ACGACCAAGGAAAGCGTCAAAAGCAAGGTGCAGTACCTGAAGGACAGGCAAGCTGGCGGGCG			1216
Query 1091	CTTTGTTTGGGCACTTGATCTGGATGACTT 1121			
Sbjct 1217	CCATGGTATGGGCCCTGGACCTGGATGACTT 1247			

Range 2: 1729 to 1740 [Graphics](#) ▼ Next Match ▲ Previous Match ▲ First

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

The screenshot shows the NCBI BLAST Standard Nucleotide BLAST interface. The interface is divided into several sections:

- Enter Query Sequence:** This section contains a text area for entering a query sequence (1), a 'Clear' button, a 'Query subrange' section with 'From' and 'To' input fields (2), an 'Or, upload file' section with a 'Choose...' button (3), and a 'Job Title' section with a text input field (4).
- Choose Search Set:** This section includes a 'Database' section with radio buttons for 'Human genomic + transcript', 'Mouse genomic + transcript', and 'Others (nr etc.):' (5), a dropdown menu for 'Nucleotide collection (nr/nt)', an 'Organism' section with a text input field and an 'Exclude' button (6), and an 'Exclude Optional' section with checkboxes for 'Models (XM/XP)' and 'Uncultured/environmental sample sequences'.
- Program Selection:** This section includes an 'Optimize for' section with radio buttons for 'Highly similar sequences (megablast)', 'More dissimilar sequences (discontiguous megablast)', and 'Somewhat similar sequences (blastn)' (7), and a 'Choose a BLAST algorithm' dropdown menu (8).
- BLAST Button:** A large blue button labeled 'BLAST' (9) is located at the bottom left of the interface.

۱. در این قسمت توالی مورد نظر را با فرمت FASTA یا شماره 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 می‌خواهیم توالی ما در آن جستجو شود. به صورت پیش فرض کلیه پایگاه‌های مشابه یا (nonredundent) nr 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 نمایی BLAST جستجو آغاز می‌شود.

Algorithm parameters

General Parameters

Max target sequences: 100
 Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask lower case letters

BLAST Search database **Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)**
 Show results in a new window


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 اندازه 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.

صفحه نتایج شامل سه بخش است:

بخش اول شامل توضیحاتی در مورد در مورد الگوریتم BLAST و نویسندگان این الگوریتم و ویرایش‌هایی است که در گذشته و حال مورد استفاده هستند. در این بخش جستجوی ما دارای یک ID است که در صورت لزوم (مثلاً بروز مشکل و درخواست کمک از مدیر پایگاه) می‌توانیم 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 می‌دهد.



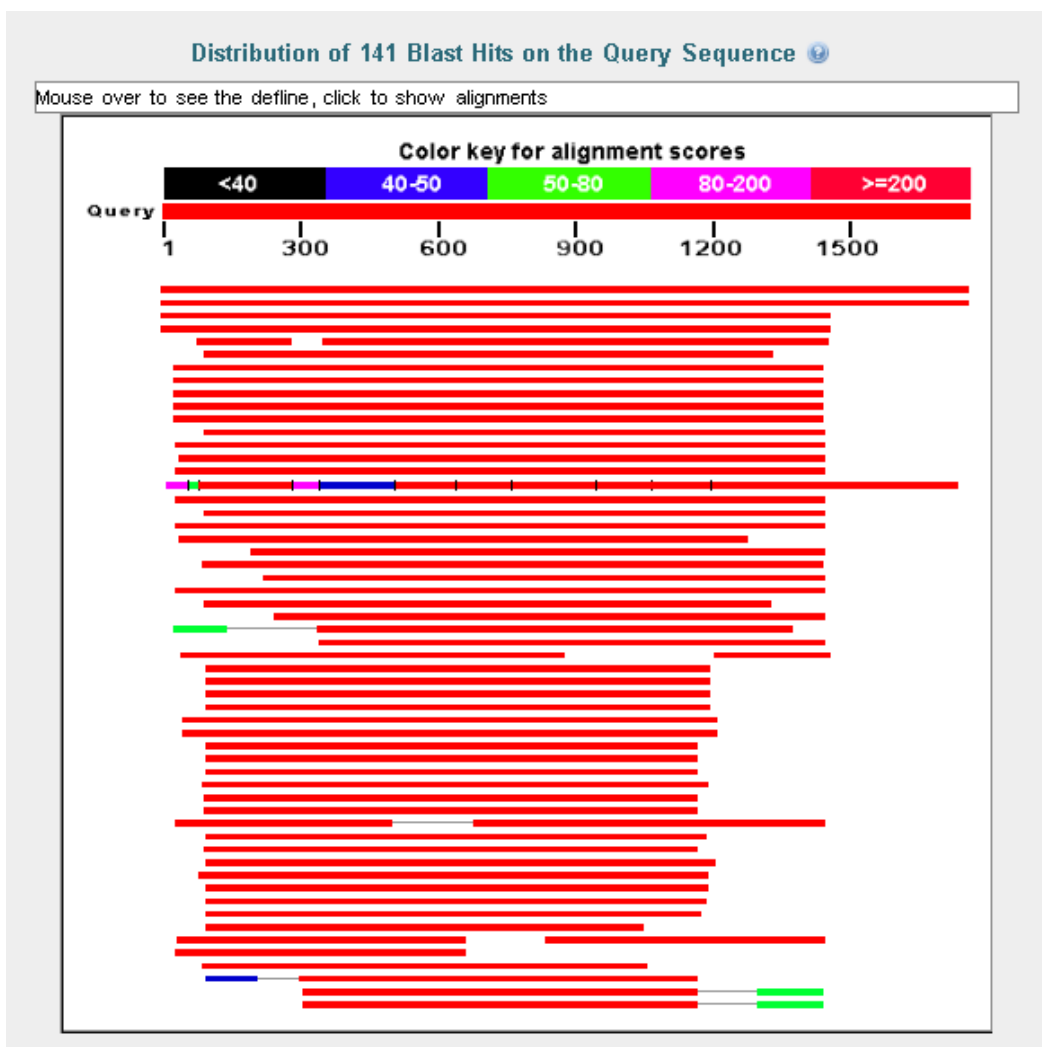
The screenshot shows the NCBI BLAST search results page. The header includes the BLAST logo and navigation links: Home, Recent Results, Saved Strategies, and Help. The main content area displays the search results for the query 'gi|47086016|ref|NM_213213.1| Danio rerio'. The results table shows the following information:

Field	Value
RID	6YTSFNHN01R (Expires on 10-30 15:43 pm)
Query ID	Id 98935
Description	gi 47086016 ref NM_213213.1 Danio rerio chitinase, acidic.3 (chia.3), mRNA
Molecule type	nucleic acid
Query Length	1746
Database Name	nr
Description	Nucleotide collection (nt)
Program	BLASTN 2.2.28+ Citation

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#)

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 توالی 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 رابطه 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 نتیجه به‌دست آمده بیشتر می‌شود. نتایج BLAST بر اساس این دو شاخص مرتب می‌شوند و بنابراین توالی‌هایی که در اوائل لیست هستند توالی‌هایی 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.

Sequences producing significant alignments:

Select: All None Selected: 0

Alignments Download GenBank Graphics Distance tree of results

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: Fundamilia nvererei 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%	CR734386.2
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%	CU571318.15
Tetraodon nigroviridis full-length cDNA	951	951	80%	0.0	75%	CR645207.2

بخش سوم نتایج Blast

در بخش چهارم جزئیات هم‌ریفی 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:

chitinase precursor [Oncorhynchus mykiss]
 Sequence ID: [ref|NP_00117855.1](#) Length: 463 Number of Matches:
[See 1 more title\(s\)](#)

Range 1: 95 to 463 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
481 bits(1239)	2e-164	Compositional matrix adjust.	255/375(68%)	309/375(82%)	6/375(1%)	+2
Query 68	VLGGWVNFGSTQFSITVSTQAMRQKFIQSSI KFLRTHGFDGLDLDWEYRGSRGSPPEDKHR					247
Sbjct 95	+GGW FG+ QFSI VS+ NR KFIQSSI FLRTH FDGLDLDWEY G+RGSPPEDK R					154
Query 248	FTLLCRELVQAYPAEAAATGKPOLMLTAAVSAAGKGTIDAGYEIAEIAKE LDFINVMTYDF					427
Sbjct 155	FTLLCKELLEAFE AEGKAVSRPRLLLTAAVAAGKGTIDSGYEIAEIAKY LDFISVMTYDF					214
Query 428	HGAWDRFTGHNSPLYRGSHDSDGLIYFNTDFAMKYWRDNGTPVEKLRVGFATYGRTRFLA					607
Sbjct 215	HG+W+ FTGHNSPLY+GSHD+GD IY NTFDFAMKYWRD G PVEKL +GFATYGR+F+LA					274
Query 608	SSNTGVgapasgaasagpYTREAGFWSYIEICTFLKGIQWIDDQKVGATKNSEWVGF					787
Sbjct 275	S ++GVGA A+GAA+AGP+TREAGFWSYIEICTFL+GAS QWIL+DQKV YA+K ++WVGF					334
Query 788	DSKESYETKVRYLQDQKYGGAFVWALDLDFFAGRCGEGSHPLLGLRKLMDVElpplpp					967
Sbjct 335	DNRESYDTKVGYLKENGFGGAMVWVLDLDDFAGQSCGQGNYPILSHLQKLLNIERPPLPP					394
Query 968	tttppkGDGQtttrpttttttttttAPGSDFCSGKADGMYANPADRNSFYVCAGGITYVRPPR					1147
Sbjct 395	T TP PG+ P T TT A GS FC+G+ADG+Y +SFY CA GIT+++					448
Query 1148	ARTVFDDSCKFCWV 1192					
Sbjct 449	A VF DSCK C WP AGLVFS DSCKCCNWP 463					

تعداد فواصل وارد شده در هم‌ردیفی

همان E-Value است که هرچقدر پایین‌تر باشد بهتر است.

تعداد جایگاه‌های یکسان در هم‌ردیفی (Identities) و جایگاه‌های یکسان به اضافه جایگاه‌های مشابه (Positives)

فریم ترجمه

توالی الگو یا توالی که ما به نرم افزار دادیم

توالی جستجو شده

توالی هم‌ردیف بین دو توالی

اعداد نشان‌دهنده شروع و اتمام بخش‌های شرکت‌کننده در هم‌ردیفی

امتیاز هم‌ردیفی بر اساس ماتریس امتیازدهی و جریمه فواصل. هرچه بالاتر باشد نشان‌دهنده شباهت بیشتر دو توالی است.

اطلاعات شناسه‌ای توالی جستجو شده یا Subject (نام توالی، طول و شماره دسترسی) آن

بخش چهارم نتایج Blast.

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	ATGGGGAGACTTACACTTATAGCAGGTTTGAGCTTGGTCTGCTGCCATGTCGCCTTTCC	87		
Sbjct 89	ATGGCCAGGCTCACAAATCTAGCAGGTTGTGCCTGGTATAAGCCAGCTGGGATCTGCC	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 هم جهت با یکدیگر هم‌ردیف شده‌اند یا به عبارت دیگر یکی روی رشته بالا (Sense) و دیگری روی رشته پایین (anti-sense) قرار گرفته است.

منبع:

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