

## فصل هفت

### توالی‌های نوکلئوتیدی، ژن‌ها و پروتئین‌ها



#### مقدمه

بررسی و مقایسه توالی‌های ژنتیکی یا توالی‌های پروتئینی کمک فراوانی به فرضیه تکامل می‌کند. به عنوان مثال، می‌توان با استفاده از نرم افزار BLAST توالی یک ژن خاص را با میلیون‌ها توالی ژنی موجود مقایسه کرد و از این طریق پی به نیای ژنتیکی بسیاری از ژن‌ها برد. بیوانفورماتیک ابزارهایی زیادی برای مطالعه بسیاری از سؤالات مربوط به حوزه بیولوژی مانند پی بردن به شباهت دو ژن خاص با عملکردهای مشابه، در اختیار محققین قرار داده است که عبارتند از تعداد زیادی پایگاه‌های داده‌ی با ارزش که حاوی اطلاعات ژن‌ها و پروتئین‌های بیان شده هستند که از بافت‌ها مشتق شده‌اند و نیز نرم افزارهایی برای تحلیل این توالی‌ها، امروزه میلیون‌ها گونه زنده وجود دارد که می‌توان آن‌ها را در سه شاخه اصلی باکتری‌ها، آرکئا (میکروب‌ها تک سلولی که در محیط‌های بی نهایت زندگی می‌کنند) و یوکاریوت‌ها گروه بندی کرد. پایگاه داده‌های مربوط به توالی‌های مولکولی هم اکنون توالی‌های DNA بیش از یکصد هزار ارگانیزم مختلف رانگهداری می‌کند. ژنوم چند صد ارگانیزم به صورت کامل تعیین توالی شده که از طریق این پایگاه‌های داده‌ی نوکلئوتیدی در دسترس است.

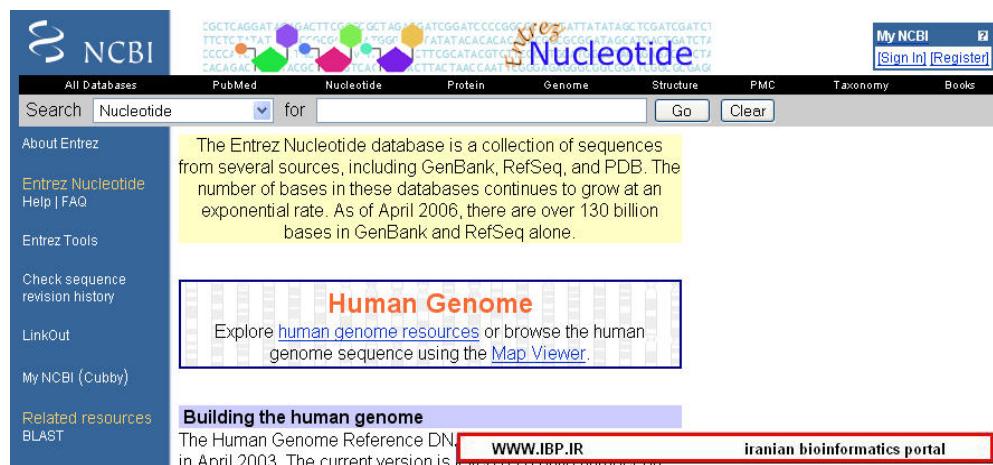
#### پایگاه داده توالی‌های نوکلئوتیدی

توالی‌های نوکلئوتیدی در سه پایگاه داده EMBL/DDBJ/Genbank نگهداری می‌شوند. داده‌های جدید هر ۲۴ ساعت بین این سه پایگاه به اشتراك گذاشته می‌شود و داده‌ها را بین هم انتقال می‌دهند. بنابراین داده‌های هر سه بانک یکسان است و جهت امنیت و حفاظت از این داده‌ها، این سه بانک در سه گوشه دنیا قرار گرفته‌اند. سازمان NCBI بانک نوکلئوتیدی بزرگ GenBank را مدیریت می‌کند. همچنان که گفته شد این سازمان در ایالت متحده قرار دارد. بانک نوکلئوتیدی

DDBJ در ژاپن و EMBL نیز در اروپا استقرار یافته‌اند. در این بخش چگونگی دستیابی به توالی‌های نوکلئوتیدی موجود در پایگاه GenBank را شرح می‌دهیم. همانند بانک‌های دیگر چندین راه برای دستیابی به این بانک نوکلئوتیدی وجود دارد. یک راه این که در صفحه اصلی NCBI جست‌جو، از منوی Search گزینه All Database را انتخاب نمایید و جست‌جویی را انجام دهید تا Entrez تمام بانک‌های شما جست‌جو نماید. سپس در صفحه نتیجه، لینک Nucleotide را انتخاب نمایید. راه ساده‌تر این که در صفحه اصلی NCBI منوی Search را باز و گزینه Nucleotide را انتخاب نمایید. در جلوی for عبارت جست‌جویی خود را تایپ نموده و دکمه Go را کلیک نمایید تا صفحه اینترنتی مطابق شکل ۱-۷ با لوگوی Nucleotide را مشاهده کنید.

### جست‌جو در ژنوم انسان

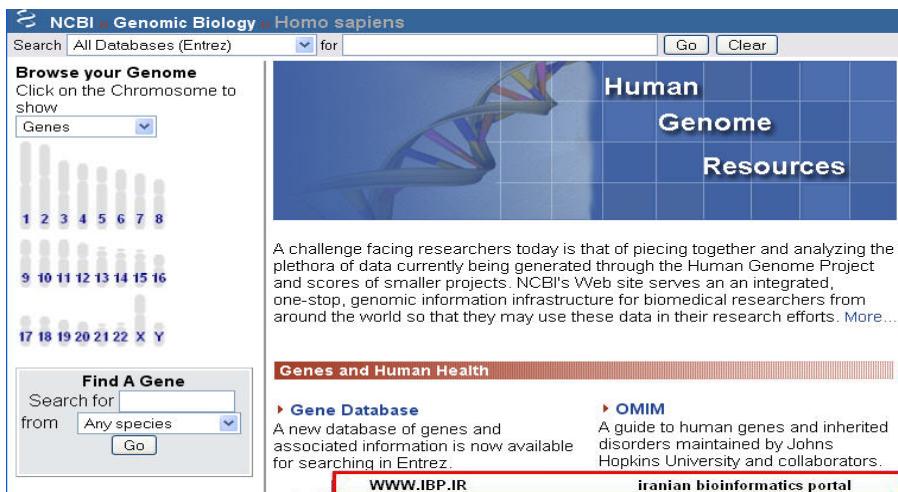
آنچه که در شکل ۱-۷ نظر شمارا به خود جلب می‌کند عبارت Genome در میان صفحه است. در این قسمت می‌توانید از نتایج پروژه بین‌المللی تعیین توالی انسان که در سال‌های ۱۹۹۶ تا ۲۰۰۳ به طول انجامید بهره‌هایی و افری ببرید.



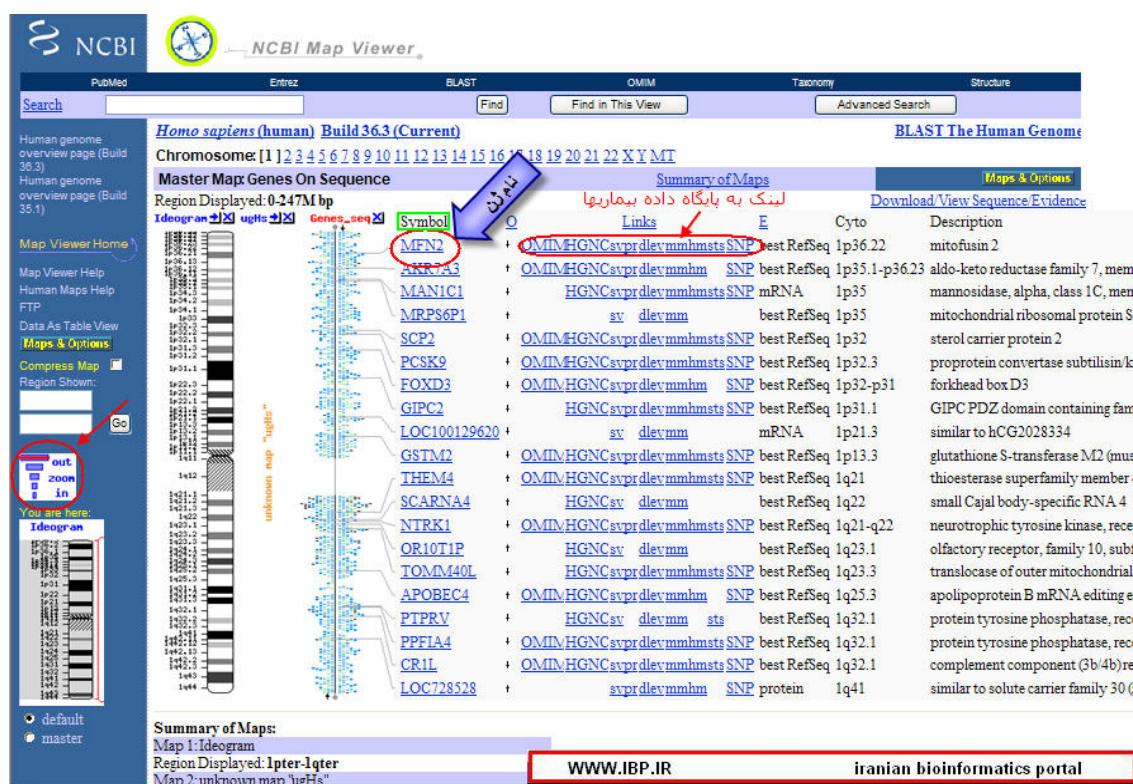
شکل ۱-۷

بر روی عبارت Human Genome Resources در میان صفحه، کلیک نمایید. صفحه زیر را مشاهده خواهید کرد. (شکل ۲-۷) در سمت چپ شکل ۲-۷ کروموزوم‌های انسان با شماره‌های آن‌ها نشان داده شده است. با کلیک بر روی هر

کدام، نقشه ژنتیکی همان کروموزوم به نمایش در می آید. به عنوان نمونه ما بر روی کروموزوم شماره یک کلیک نموده ایم و به صفحه ای مطابق شکل ۳-۷ دست یافته ایم.



شکل ۲-۷



شکل ۳-۷

این صفحه بسیار شلوغ به نظر می رسد ولی برای محققینی که در زمینه نقشه های ژنتیکی کار کرده اند اطلاعات ذی قیمتی دارد. در این تصویر، نوار بندی

کروموزوم شماره یک را می بینید. بر روی این کروموزوم، نام ژن‌ها و بسیاری اطلاعات مفید دیگر به چشم می خورد.

اگر بر روی نواربندی کروموزوم کلیک نمایید آن منطقه از کروموزوم را برای شما بزرگتر به نمایش در می آورد. در ستون Symbol نام ژن‌های روی این کروموزوم را نشان می دهد. و در سمت راست این ستون، هم ردیف با بعضی از ژن‌ها کلمه OMIM را مشاهده می نمایید که لینک به پایگاه داده بیماری‌های ژنتیکی است. می توانید بر روی آن کلیک نمایید تا به آن پایگاه داده دست یابید.

### جست و جو در بانک نوکلئوتیدی‌ها

اکنون دوباره به بحث اصلی خود که جست و جو در پایگاه داده Nucleotide بود بر می گردیم جست و جو خود را در این بانک شروع می کنیم. به صفحه اینترنتی مطابق شکل ۱-۷۱ برگردید و عبارت جست و جوی خود را در قسمت Search Box وارد و دکمه Go را کلیک کنید. عبارت جست و جوی ما در اینجا همان عبارتی است که در بانک‌های دیگر نیز از آن استفاده کردیم یعنی:

DCC AND "Vogelstein B" [AU]

هدف ما یافتن توالی نوکلئوتیدی ژن DCC است که در تحقیقات خود از آن استفاده نماییم. بعد از جست و جو نتیجه را در شکل (۴-۷ a) مشاهده می نمایید.

Found 17 nucleotide sequences.

	CoreNucleotide records
0	EST (Expressed Sequence Tags) records
0	GSS (Genomic Survey Sequence) records

**What is this?**

The Nucleotide database has been recently merged into three databases:

- CoreNucleotide contains all Nucleotide records that are not in EST or GSS. They are of interest to most users.

[WWW.IBP.IR](http://WWW.IBP.IR) [iranian bioinformatics portal](#)

شکل ۴-۷a

Search CoreNucleotide for DCC AND "Vogelstein B"[AU] Go Clear Save Search

Display Summary Show 20 Sort by Send to

All: 17 Bacteria: 0 RefSeq: 2 mRNA: 4

Items 1 - 17 of 17

One page. Top Organisms [Tree]

**1:** NM\_007831 Reports  
Mus musculus deleted in colorectal carcinoma (Dcc), mRNA  
gi|133778956|ref|NM\_007831.3|[133778956]  
اطلاعات در مورد موش

**2:** NM\_005215 Reports  
Homo sapiens deleted in colorectal carcinoma (DCC), mRNA  
gi|110431347|ref|NM\_005215.2|[110431347]

**3:** X76132 Reports  
H.sapiens DCC mRNA  
gi|453209|emb|X76132.1|[453209]

by:babak babaabasi  
Links babakbabaabasi@gmail.com

WWW.IBP.IR iranian bioinformatics portal

شكل ۴-۷b

همان طور که ملاحظه می کنید تعداد ۱۷ رکورد یافت شده است که با شماره از هم جدا شده اند. ساختار این صفحه مقاومت از ساختار جست و جوی بانک های دیگر است. نگران نباشید؛ ما همه را مفصلابرا برای شما شرح خواهیم داد.  
درباره هر رکورد اطلاعات زیر وجود دارد:

- در سطر اول در کنار هر شماره عبارتی را می بینید که با NM شروع شده است. این عبارت کد یکتای رکورد است که به آن LOCUS ID گفته می شود. تنها با این کد می توانید به این رکورد دست یابید بدین ترتیب که این کد را به عنوان عبارت جست و جوی خود در Search Box وارد نمایید. بنابر این اگر در جست وجوها، رکوردي را مناسب تحقیقات خود دانستید کافی است که این ID را یادداشت نمایید.

- در سطر دوم، عنوان این رکورد قرار گرفته است. با استفاده از این عنوان می توانید پی به اطلاعات مهمی درباره این رکورد ببرید. مثلا در رکورد شماره دوم عبارت Homo sapiens در ابتدای عنوان رکورد بیان می کند که اطلاعات ژنتیکی این رکورد مربوط به انسان است. یا در رکورد اول Mus musculus اشاره می کند که اطلاعات رکورد در مورد موش است.

- در سطر سوم عدد بعد از gi، کد این رکورد در بانک GenBank است.
  - اگر بر روی لینک هر رکورد (مثلاً در رکورد اول NM\_007831) که در سطر اول قرار گرفته است، کلیک کنید جزئیات هر رکورد همراه با توالی نوکلئوتیدی آن مطابق شکل ۵-۷ نمایش داده می‌شود که ما در ادامه همه را کاملاً
- شرح می‌دهیم.

□1: NM\_007831. Reports Mus musculus dele..[gi:133778956] Links

[Comment](#) [Features](#) [Sequence](#)

**LOCUS** NM\_007831 10325 bp mRNA linear ROD 16-MAR-2008  
**DEFINITION** Mus musculus deleted in colorectal carcinoma (Dcc), mRNA.  
**ACCESSION** NM\_007831  
**VERSION** NM\_007831.3 GI:133778956  
**KEYWORDS** .  
**SOURCE** Mus musculus (house mouse)  
**ORGANISM** Mus musculus

[WWW.IBP.IR](http://WWW.IBP.IR) iranian bioinformatics portal

شکل ۵-۷

### بررسی یک رکورد نمونه

محتوای این صفحه (شکل ۵-۷) زیاد است. برای شرح جزئیات این صفحه، نمونه ای کوچک‌تر را در صفحه بعد ارایه کرده‌ایم و در پایان این رکورد نیز توضیحات کاملی درباره آن داده‌ایم.

در نگاه اول، این اطلاعات بغرنج و نامفهوم به نظر می‌رسند. جزء جزء این اطلاعات را مورد بررسی قرار داده و مفهوم آن‌ها را شرح می‌دهیم. به این ساختار ارایه شده fault file گفته می‌شود. اطلاعات درون fault file ها به صورت Line type ارایه می‌شود. بدین معنا که در اول هر سطر، یک کلمه یا یک کد قرار گرفته است و بیان می‌کند در آن سطر چه اطلاعاتی گنجانده شده است. برای مثال در سطر اول کلمه LOCUS در ابتدای سطر، بیان می‌کند که در این سطر کد یکتاً رکورد (SCU49845) یا Locus ID (SCU49845) قرار دارد. یا در سطري که در ابتدای آن

کلمه ORGANIIM آمده، بیان کننده درخت تاکسonomic است که این ژن از آن گرفته شده است.

```

LOCUS          SCU49845                      5028 bp      DNA      linear      PLN
21-JUN-1999
DEFINITION    Saccharomyces cerevisiae TCP1-beta gene, partial cds; and
Ax12p        (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION     U49845
VERSION       U49845.1  GI:1293613
KEYWORDS      .
SOURCE         Saccharomyces cerevisiae (baker's yeast)
ORGANISM      Saccharomyces cerevisiae
Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
Saccharomyces.
REFERENCE     1  (bases 1 to 5028)
AUTHORS       Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE         Cloning and sequence of REV7, a gene whose function is
required for DNA damage-induced mutagenesis in Saccharomyces
cerevisiae
JOURNAL      Yeast 10 (11), 1503-1509 (1994)
PUBMED       7871890
REFERENCE     2  (bases 1 to 5028)
AUTHORS       Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE         Selection of axial growth sites in yeast requires Ax12p,
a novel plasma membrane glycoprotein
JOURNAL      Genes Dev. 10 (7), 777-793 (1996)
PUBMED       8846915
REFERENCE     3  (bases 1 to 5028)
AUTHORS       Roemer,T.
TITLE         Direct Submission
JOURNAL      Submitted (22-FEB-1996) Terry Roemer, Biology, Yale
University, New Haven, CT, USA
FEATURES      Location/Qualifiers
source        1..5028
              /organism="Saccharomyces cerevisiae"
              /mol_type="genomic DNA"
              /db_xref="taxon:4932"
              /chromosome="IX"
              /map="9"
              <1..206
              /codon_start=3
              /product="TCP1-beta"
              /protein_id="AAA98665.1"
              /db_xref="GI:1293614"
CDS           /translation="SSIYNGISTSGLDLNNGTIADMRLGIVESYKLKRAVVSSASEA
AEVLLRVDNIIIRARPRTANRQHM"
gene          687..3158
              /gene="AXL2"
CDS           687..3158
              /gene="AXL2"
              /note="plasma membrane glycoprotein"

```

```

/codon_start=1
/product="Ax12p"
/protein_id="AAA98666.1"
/db_xref="GI:1293615"

gene complement( 3300 .. 4037 )
/gene="REV7"
CDS complement( 3300 .. 4037 )
/gene="REV7"
/codon_start=1
/product="Rev7p"
/protein_id="AAA98667.1"
/db_xref="GI:1293616"



```

301 agacgcgaaa aaaaaagaac aacgcgtcat agaacctttg gcaattcgcg  
 tcacaataa  
 361 attttggcaa cttatgttcc ctcttcgagc agtactcgag ccctgtctca  
 agaatgtaat  
 421 aatacccatc gtaggtatgg ttaaaagatag catctccaca acctcaaagc  
 tccttgccga  
 481 gagtcgcccct cctttgtcga gtaatttca cttttcatat gagaacttat  
 tttcttattc  
 541 tttaactctca catcctgttag tgattgacac tgcaacagcc accatcacta  
 gaagaacaga  
 601 acaattactt aatagaaaaaa ttatatcttc ctcgaaacga tttcctgctt  
 ccaacatcta  
 661 cgtatatcaa gaagcattca cttaccatga cacagctca gatttcattta  
 ttgctgacag  
 721 ctactatatac actactccat ctagtagtgg ccacgcccata tgaggcataat  
 cctatcgaa  
 781 aacaataccc cccagtggca agagtcaatg aatcgttac atttcaaatt  
 tccaatgata  
 841 cctataaaatc gtctgttagac aagacagctc aaataacata caattgcttc  
 gacttaccga  
 901 gctggcttcc gtttgactct agttctagaa cgttctcagg tgaaccttct  
 tctgacttac  
 961 tatctgtatgc gaacaccacg ttgtatattca atgtaataact cgagggtacg  
 gactctgccg  
 1021 acagcacgta tttgaacaat acataccaat ttgttgttac aaaccgtcca  
 tccatctcgc  
 1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttataact  
 aacggcaaaa  
 1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt  
 tcaatgttca  
 1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat  
 gcgccgttac  
 1261 ccaattggct gttcttcgat tctggcgagt tgaagttac tgggacggca  
 ccggtgataa  
 1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca  
 gacattgaag  
 1381 gattttctgc cggtgaggta gaattcgaat tagtcattcgg ggctcaccag  
 ttaactacct  
 1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca  
 tatgacttac  
 1501 ctctaaacta tgtttatctc gatgacgatc ctatttcttc tgataaatttgc  
 ggttctataa  
 1561 acttatttggta tgctccagac tgggtggcat tagataatgc taccatttcc  
 gggtctgtcc  
 1621 cagatgaatt actcggtaag aactccaatc ctgccaattt ttctgtgtcc  
 attttagata  
 1681 cttatggta tgtgattttat ttcaacttcg  
 //  
 به علت طولانی بودن توالی قسمتی از انتهای توالی را حذف کردیم تا تعداد صفحات زیاد گیج کننده نباشد

## کالبد شکافی Flat file های توالی‌های نوکلئوتیدی

Flat file ها به سه بخش اصلی تقسیم می‌شوند:

حاوی اطلاعاتی (descriptors) است در مورد کل رکوردها (۱)

(ویژگی‌ها) که یادداشت‌ها و توضیحات رکورد است. features (۲)

(۳) و خود توالی نوکلئوتیدی.

همه flat file های پایگاههای نوکلئوتیدی دارای علامت // در انتهای هر رکورد هستند. اکنون هر سه بخش بالا را مفصل شرح می دهیم.

### The header

بخش header بخشی است که بیش ترین وابستگی را به پایگاه داده دارد. یعنی این که در پایگاه دادههای نوکلئوتیدی مختلف تقاضه هایی مشاهده می شود. سطر اول همه flat file ها در بانک GenBank سطر LOCUS است که در زیر نمونه ای از آن را می بینید:

LOCUS SCU 49845 5028 bp DNA PLN 21-JUN-1999

• در این سطر، ابتدا یک نام دلخواه (SCU49845) می آید که به آن، نام رکورد گفته می شود. این المان (نام) باید با یک حرف شروع شود و کاراکتر بعدی می تواند حرف یا عدد باشد. در این مثال سه کاراکتر اول، حرف و بقیه عدد است. همه حروف نام همگی حرف بزرگ هستند. طول این المان باید از ۱۰ کاراکتر بیش تر نشود. در گذشته افرادی که توالی های درون بانک را کنترل می کردند، می کوشیدند تا نام مفید و با مسما انتخاب کنند و در ضمن این که درون بانک را کنترل می کردند، می کوشیدند تا نام مفید و با مسما انتخاب کنند و در ضمن این که این نام می بایست در کل بانک یکتا باشد، از آن جا که تمام نامهای با معنا تاکنون استفاده شده است، نامهای جدید معنای خاصی ندارند. امروزه این نام، تنها از این نظر در بانک حضور دارند که نرم افزارهای قدیمی زیادی وابسته به این فیلد هستند. بنابراین این نامها حاوی اطلاعات مفیدی نیستند.  
• دومین المان در خط LOCUS، طول توالی است. در این مثال طول توالی 5028 bp است.

• سومین المان در این سطر نشان دهنده ماهیت بیولوژی مولکول است که معمولاً RNA یا DNA است.

• المان چهارم در این سطر (PLN)، کد تقسیم بندی (division code)، کد تقسیم بندی (division code) است. این کد سه حرفی، کد تقسیم بندی تاکسانومی بوده یا برای دسته بندی های دیگری به کار می رفته است، NCBI کد تقسیم بندی تاکسانومی ارگانیزمها را

به روز نکرده است، چرا که فکر می کرد این کد ۳ حرفی مناسب نمایش گونه های متنوع بیولوژی موجود روی کره زمین نیست. این نوع کنگذاری امروزه منسخ شده است. جدیدا این تقسیم بندی بر پایه عملکرد (function) (ژن ها صورت می گیرد. ممکن است به جای این کدها با عباراتی نظیر EST یا STS برخورد نمایید. توضیح مختصری از این دو کد در زیر آورده شده است:

## EST

این کلمه مخفف Expressed Sequence است. این تقسیم بندی در سال ۱۹۹۳ معرفی شده است. رکوردهایی که دارای این کد سه حرفی هستند، حاوی توالي های کوتاه از cDNA هستند. معمولاً به تعداد زیادی از روی این نوع ژن رونویسي صورت می گیرد. EST ها نمایش یک تصویر فوري از آنچه در یک بافت خاص یا در یک مرحله خاصی از رشد موجود (development) بیان می شود. این ها tag هایی از بین یک ژن خاص موجود در یک کتابخانه خاص cDNA هستند.

## STS

این کلمه مخفف "Sequence Tagged Site" است. STS ها توالي های کوتاه (بین ۲۰۰ الی ۵۰۰ جفت باز) و روی ژنوم یکتا هستند. جایگاه و توالي آن ها کاملا مشخص است. این توالي ها قابل آشکار شدن به وسیله PCR هستند. توالي های STS، به دلیل یکتا بودن روی ژنوم، از آن ها جهت تهیه نقشه های ژنتیکی استفاده می شود.  
• المان پنجم تاریخ اضافه شدن این رکورد به GenBank است. در این مثال: 1998-22-FEB. اگر هر یک از ویژگی های رکورد به روز شده باشد و سپس رکورد انتشار یابد، این تاریخ، آخرین تاریخ انتشار است.  
بخش بعدی definition، سطر header به فرمت زیر است:

DEFINITION Saccharomyces cereviiae TCP1-beta gane , partial  
cds, and Ax12p(AXL2) and Rev7p (REV7) genes, completed cds.

در این سطر، اطلاعاتی ارایه می شود که نشان دهد این رکورد راجع به چه موضوعی است. در ایجاد این خطوط دقت فراوانی صورت گرفته است. گرچه

بسیاری از این اطلاعات را می توان از بخش های دیگر رکورد به دست آورد ولی این اطلاعات چنان انتخاب شده اند تا دارای ارزش باشند. بنابراین در مراجعه به این رکورد سری به این سطر بزنید تا مطمئن شوید که اطلاعات رکورد مناسب تحقیقات شما است یا خیر. به هر حال همیشه امکان ندارد که تمام اطلاعات بیولوژیک رکورد را بتوان در یک سطر جمع آوری کرد. کاربران باید تنها به این سطر اکتفا نکنند.

به پایین‌تر رکورد که بیابیم، کد دسترسی (ACCESSION)، کلید اصلی برای ارجاع به رکورد در پایگاه داده را نشان می دهد. مانند PMID در پایگاه داده PubMed، کد دسترسی همیشه با رکورد خواهد بود و اگر رکورد تغییر کند (مثلاً با تغییر یک نوکلئوتید) کد دسترسی تغییر نخواهد کرد. هم اکنون کد دسترسی به یکی از شکل های رو به رو است: "۱+۵" و "۲+۶"

"۱+۵" یعنی یک حرف بزرگ انگلیسی در اول و پنج رقم به دنبال آن، در مثال بالا کد دسترسی U49845 است که از این قانون پیروی می کند.

"۲+۶" یعنی دو حرف بزرگ انگلیسی در ابتدا و شش رقم به دنبال آن، اکثر رکوردهای جدیدی که امروزه وارد بانک می شوند کدی به صورت این فرمت دوم دارند.

سطر بعدی در این رکورد به VERSION تعلق دارد:

VERSION U49845.1 GI: 1293613

قالب version به صورت accession. version است. (U49845.1). شماره ثابت بوده ولی version هر گاه که توالی تغییر کند یک واحد به آن افزوده می شود. سطر VERSION در GenBank همچنین دارای عدد gi (the geninfo identifier) است. که در این مثال 1293613: GI: است. اگر توالی تغییر کند gi به عدد صحیح موجود بعدی تغییر می کند. عدد gi فقط مخصوص GenBank است. عدد accession. version به عنوان نام اصلی رکورد (identifier) است که تنها و تنها به یک رکورد در هر بانک اشاره می کند.

سطر بعدی حاوی کلمات کلیدی (keywords) رکورد است. این خطوط به دلایل تاریخی وجود دارند و در بسیاری از موارد متاسفانه بی استفاده‌اند.

اطلاعات Taxonomy در سطرهایی با عنوان SOURCE و ORGANISM در GenBank وجود دارد. این کلاس بندی از بالا به پایین است. مطابق درخت taxonomic عمومی‌ترین و بزرگترین گروه، اول آورده می‌شود. در این مثال اطلاعات تاکسonomی موجودی که اطلاعات ژنتیکی آن در این رکورد قرار دارد به صورت زیر است:

<b>SOURCE</b>	<i>Saccharomyces cerevisiae (baker's yeast)</i>
<b>ORGANISM</b>	<u><i>Saccharomyces cerevisiae</i></u> <i>Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;</i> <i>Saccharomycetes; Saccharomycetales; Saccharomycetaceae;</i> <i>Saccharomyces.</i>

هر رکورد حداقل یک مرجع (citation) یا (reference) داشته باشد. این مراجعها مقالاتی هستند که به توالی اعتبار می‌دهند که چرا این ژن خاص تعیین توالی شده است. در بسیاری از موارد رکوردها دو یا چند مرجع دارند. نمونه‌ای از یک مرجع در زیر آمده است:

<b>REFERENCE</b>	1 (bases 1 to 5028)
<b>AUTHORS</b>	Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
<b>TITLE</b>	Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in <i>Saccharomyces cerevisiae</i>
<b>JOURNAL</b>	<i>Yeast</i> 10 (11), 1503-1509 (1994)
<b>PUBMED</b>	<a href="#">7871890</a>

اگر به رکورد اصلی مراجعه نمایید متوجه می‌شوید که سه مرجع با ساختار بالا وجود دارد. مراجعها با شماره‌ای که دقیقاً جلوی کلمه REFERENCE قرار گرفته است از هم جدا شده‌اند. افزودن یک رکورد به بانک برای فرد اعتبار علمی ایجاد می‌کند. هر مراجعت از بخش‌های زیر تشکیل شده است.

•نویسنگان مقاله (AUTHORS)

•عنوان مقاله (TITLE)

•نام ژورنال و سال انتشار آن (JOURNAL)

•کد PMID این مقاله (PUBMED)

از طریق این مراجع بین دو بانک نوکلئوتیدها و بانک PubMed ارتباط برقرار می شود. بدین ترتیب می توانید با استفاده از PMID این مرجع ها به مقالات مرتبط با این رکورد دست یابید.

## The Feature Table

بخش میانی رکورد، FEATURES است. در قسمت FEATURES اطلاعات ژنتیکی این ژن DCC که آن را مورد جستجو قرار داده ایم) قرار دارد. این بخش از رکورد در زیر آمده است.

<i>FEATURES</i>	<i>Location/Qualifiers</i>
<i>source</i>	1..5028 /organism="Saccharomyces cerevisiae" /mol_type="genomic DNA" /db_xref="taxon: <a href="#">4932</a> " /chromosome="IX" /map="9" <i>CDS</i> <1..206 /codon_start=3 /product="TCP1-beta" /protein_id=" <a href="#">AAA98665.1</a> " /db_xref="GI:1293614"
<i>translation</i>	<code>SSIYNGISTSGLDLNNGTIADMRLGIVESYKLKRAVVSSASEA AEVLLRVDNIIIRARPRTANRQHM"</code>
<i>gene</i>	687..3158 /gene="AXL2"

اطلاعات موجود در این قسمت دسته بندی شده است. دسته ای از این اطلاعات در قسمت source قرار دارد که به کل ژن اشاره دارد و دسته ای دیگر در قسمت CDS و بخشی نیز در قسمت gene قرار دارد.

تفسیر این اطلاعات به صورت زیر است:

- طول این ژن ۵۰۲۸ نوکلئوتید است که در جلوی source به صورت ۱...5028 نشان داده شده است.

- این ژن مربوط به "ساکارومایسیس سروسیه" است که در سطر source /organism در بخش قرار گرفته است.

- کد تاکسانومی این موجود در سطر db\_xref، عدد ۴۹۳۲ را نشان می دهد.

- این ژن روی کروموزوم IX این موجود قرار دارد که در بخش source/در زیر chromosome مشاهده می کنید.
- این ژن در قسمت ۹ ام کروموزوم IX قرار دارد که این اطلاعات در بخش map/زیر source می بینید.
- توالی این ژن از سه exon تشکیل شده است که اطلاعات هر کدام در یک CDS جدگانه گذاشته شده است. همچنان که در قسمت اولین CDS می بینید اولین اگزون از باز ۱ شروع و به باز ۲۰۶ ختم می شود. کدون شروع آن باز شماره ۳ است (codon\_start=3). نام این پروتئین TCP1-beta، شناسه آن AAA98665 در سطر های بعدی آمده است. ترجمه این exon نیز در سطر translation مشاهده می نمایید.
- نام این ژن AXL2، در سطر gene قرار دارد.

## توالی ژن‌ها

در آخر رکورد نیز در بخش ORIGIN توالی کامل این ژن وجود دارد. نوکلئوتیدهای این ژن در پنج دسته ده تایی در هر سطر قرار دارد که در این نمونه به دلیل کمبود فضا آن را در چهار دسته قرار داده‌ایم. در بسیاری از نرم افزارهای کامپیوتری از این توالی نوکلئوتیدی به عنوان ورودی نرم افزار استفاده می شود مانند نرم افزار BLAST که توالی‌های نوکلئوتیدی را با یکدیگر مقایسه می نماید.

نکته آخر در این فصل :

احتمالاً در بسیاری از مطالب مربوط با نرم افزارهای بیوانفورماتیکی به کلمه FASTA برخورد کرده اید یا خواهید کرد این نکته را به خاطر بسپارید که FASTA یک نوع فرمت میباشد مثلاً توی عکسها کامپیوتر و یا موزیک های کامپیوتري با فرمتهای زيادي آشنا هستيد

در داده های بیوانفورماتیکی فرمتهای خاصی برای ذخیره کردن اطلاعات وجود دارد.

در بالای صفحه اي که توالی ژن را بررسی کردیم اگر توجه کرده باشید بخش Display را دیده اید که برای تغیير دادها به فرمتهای مختلف میباشد

نکته آخر در این فصل :

احتمالاً در بسیاری از مطالب مربوط با نرم افزارهای بیوانفورماتیکی به کلمه FASTA برخورد کرده اید یا خواهید کرد این نکته را به خاطر بسپارید که FASTA یک نوع فرمت میباشد مثلاً توی عکسها کامپیوتر و یا موزیک های کامپیوتري با فرمتهای زيادي آشنا هستيد

در داده های بیوانفورماتیکی فرمتهای خاصی برای ذخیره کردن اطلاعات وجود دارد.

در بالای صفحه اي که توالی ژن را بررسی کردیم اگر توجه کرده باشید بخش Display را دیده اید که برای تغیير دادها به فرمتهای مختلف میباشد

شکل ۶-۷

مشاهده میکنید که با انتخاب فرمت FASTA

اطلاعات و توضیحات بالای توالی حذف شده و توالی زنوم بدون دسته بندی داده اند

که حال متوانید از اطلاعات به این فرمت در نرم افزارهای بیوانفورماتیکی استفاده کنید

اطلاعات و توالی زنوم به صورت مخصوص که در نرم افزارهای توالی مخفظت میگردند

جای این خط یک کلمه که در نرم افزارهای توالی مخفظت میگردند

WWW.IBP.IR      iranian bioinformatics portal

شکل ۷-۷

منابع:

۱. ابزارهای بیوانفورماتیک نویسنده‌گان: فاطمه هنری و حبیب اسلام زاده

برای اطلاعات بیشتر در رابطه با این کتاب میتوانید به آدرس [www.ibp.ir](http://www.ibp.ir) مراجعه کنید  
پیشنهاد میکنم این کتاب را تهیه کرده و به طور کامل مطالعه کنید

۲. مطالب موجود در سایت پرتال بیوانفورماتیک ایرانیان از نوشه‌های بابک باباعباسی

منابع زیاد دیگری در تهیه این کتاب الکترونیک استفاده شده است که در بخش  
معرفی کتب پرتال بیوانفورماتیک ایرانیان میتوانید مطالعه کنید [www.ibp.ir](http://www.ibp.ir)

\*\*\*\*\*