

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 ابزارهای 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 داده‌ای 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 را نگهداری می‌کنند. ژنوم چند صد ارگانیزم به صورت کامل تعیین توالی شده که از طریق این پایگاه‌های داده‌ای نوکلئوتیدی در دسترس است.

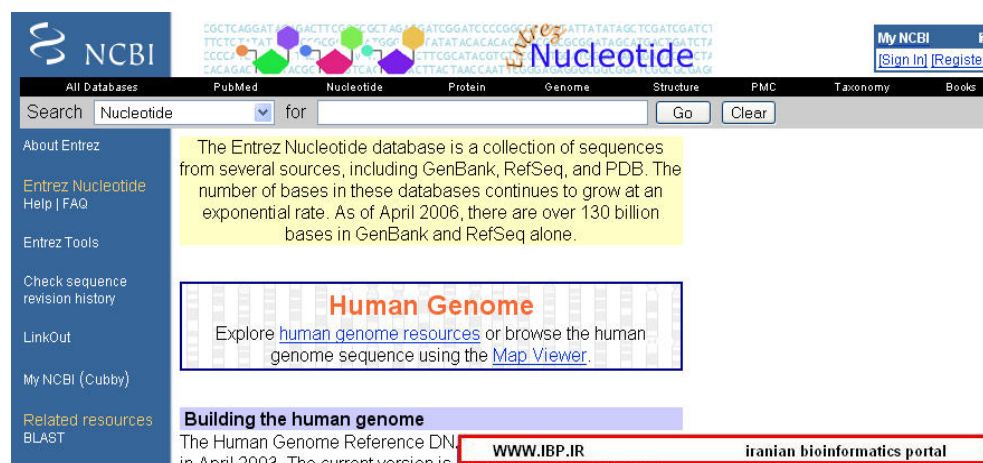
پایگاه داده توالی‌های نوکلئوتیدی

توالی‌های نوکلئوتیدی در سه پایگاه داده EMBL/DDBJ/Genbank نگهداری می‌شوند. داده‌های جدید هر ۲۴ ساعت بین این سه پایگاه به اشتراک 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. يك راه این که در صفحه اصلی NCBI جهت جست و جو، از منوی Search گزینه All Database را انتخاب نمایید و جست و جویی را انجام دهید تا Entrez تمام بانک‌ها را برای شما جست و جو 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 عبارت جست و جوی خود را تایپ نموده و دکمه Go را کلیک نمایید تا صفحه اینترنتی مطابق شکل ۱-۷ با لوگوی Nucleotide را مشاهده کنید.

جست و جو در ژنوم انسان

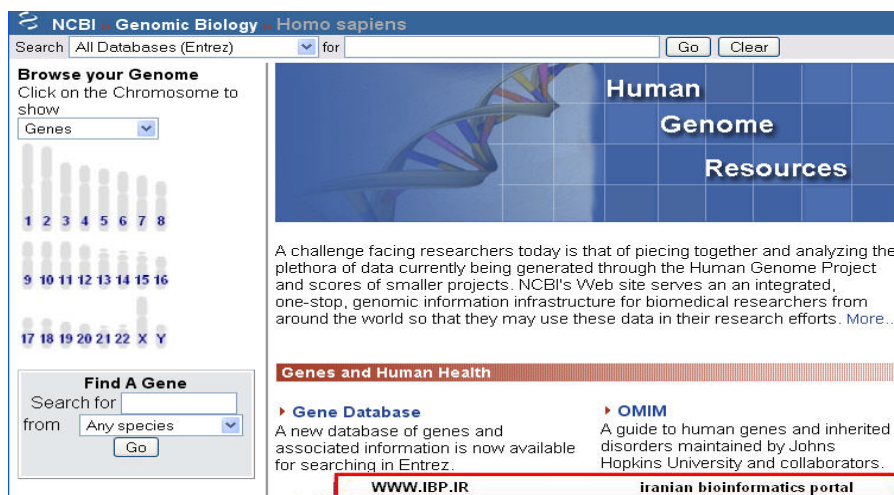
آنچه که در شکل ۱-۷ نظر شما را به خود جلب می‌کند عبارت Genome Human در میان صفحه است. در این قسمت می‌توانید از نتایج پروژه بین‌المللی تعیین توالی انسان که در سال‌های ۱۹۹۶ تا سال ۲۰۰۳ به طول انجامید بهره‌های و آفری ببرید.



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 ۷-۳ دست یافته ایم.



شکل ۷-۳

Master Map: Genes On Sequence
Region Displayed: 0-247M bp

Symbol	Links	E	Cyto	Description	
MFN2	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1p36.22	mitofusin 2
AKR1A3	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1p35.1-p36.23	aldo-keto reductase family 7, mem1
MAN1C1	HGNC:svpr:dvmm:hmsts	SNP	mRNA	1p35	mannosidase, alpha, class 1C, mem
MRPS6P1	svpr:dvmm		best RefSeq	1p35	mitochondrial ribosomal protein S6
SCP2	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1p32	sterol carrier protein 2
PCSK9	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1p32.3	proprotein convertase subtilisin/ke
FOXD3	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1p32-p31	forkhead box D3
GIPC2	HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1p31.1	GIPC PDZ domain containing fami
LOC100129620	svpr:dvmm		mRNA	1p21.3	similar to hCG2028334
GSTM2	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1p13.3	glutathione S-transferase M2 (mus
THEM4	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1q21	thioesterase superfamily member 4
SCARNA4	HGNC:svpr:dvmm		best RefSeq	1q22	small Cajal body-specific RNA 4
NTRK1	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1q21-q22	neurotrophic tyrosine kinase, recey
OR10T1P	HGNC:svpr:dvmm		best RefSeq	1q23.1	olfactory receptor, family 10, subf
TOMM40L	HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1q23.3	translocase of outer mitochondrial
APOBEC4	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1q25.3	apolipoprotein B mRNA editing er
PTPRV	HGNC:svpr:dvmm:st		best RefSeq	1q32.1	protein tyrosine phosphatase, rece
PPFIA4	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1q32.1	protein tyrosine phosphatase, rece
CR1L	OMIM:HGNC:svpr:dvmm:hmsts	SNP	best RefSeq	1q32.1	complement component (3b/4b)rec
LOC728528	svpr:dvmm:hmsts	SNP	protein	1q41	similar to solute carrier family 30 (z

Summary of Maps:
Map 1: Ideogram
Region Displayed: 1pter-1qter
Map 2: unknown map "hg18"

شکل ۷-۳

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

کروموزوم شماره يك را مي بينيد. بر روي اين کروموزوم، نام ژن ها و بسياري اطلاعات مفيد ديگر به چشم مي خورد.

اگر بر روي نواربندي کروموزوم كليك نماييد آن منطقه از کروموزوم را براي شما بزرگتر به نمايش در مي آورد. در ستون Symbol نام ژن هاي روي اين کروموزوم را نشان مي دهد. و در سمت راست اين ستون، هم رديف با بعضي از ژن ها کلمه OMIM را مشاهده مي نماييد که لينک به پايگاه داده بيماري هاي ژنتيكي است. مي توانيد بر روي آن كليك نماييد تا به آن پايگاه داده دست يابيد.

جست و جو در بانک نوکلوتیدی ها

اکنون دوباره به بحث اصلي خود که جست و جو در پایگاه داده Nucleotide بود بر مي گرديم جست و جو خود را در اين بانک شروع مي کنيم. به صفحه اينترنتي مطابق شکل ۷-۱ برگريد و عبارت جست و جوي خود را در قسمت Search Box وارد و دکمه Go را كليك کنيد. عبارت جست و جوي ما در اين جا همان عبارتي است که در بانک هاي ديگر نيز از آن استفاده کردیم يعني:

DCC AND "Vogelstein B" [AU]

هدف ما يافتن توالي نوکلوتيدي ژن DCC است که در تحقيقات خود از آن استفاده نماييم. بعد از جست و جو نتيجه را در شکل (a ۷-۴) مشاهده مي نماييد.

NCBI Nucleotide search results for "DCC AND \"Vogelstein B\" [AU]". Found 17 Nucleotide sequences.

Record Type	Count
CoreNucleotide records	17
EST (Expressed Sequence Tags) records	0
GSS (Genomic Survey Sequence) records	0

What is this?
The Nucleotide database has been recently updated with three databases:

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

WWW.IBP.IR | iranain bioinformatics portal

شکل ۷a-۴

NCBI Nucleotide

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

Limits Preview/Index History Clipboard Details

Found 17 nucleotide sequences. CoreNucleotide [17]

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]

Homo sapiens (12)
Rattus norvegicus (4)
Mus musculus (1)

Links

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]

Links

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

شکل ۷b- ۴

همان طور که ملاحظه می کنید تعداد ۱۷ رکورد یافت شده است که با شماره از هم جدا شده اند. ساختار این صفحه متفاوت از ساختار جست و جوی بانک های دیگر است. نگران نباشید؛ ما همه را مفصلاً 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. بنابراین اگر در جست و جوها، رکوردی 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. مثلاً 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 جزییات هر رکورد همراه با توالی نوکلئوتیدی آن مطابق شکل ۵-۷ نمایش داده می شود که ما در ادامه همه را کاملاً شرح می دهیم.

NCBI Nucleotide

Search: CoreNucleotide for scu49845

Display: GenBank Show 5 Send to Hide: ☐ sequence ☐ all but gene, CDS and mRNA features

Range: from begin to end ☐ Reverse complemented strand Features: ☒ STS ☒ Exon

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 iranain bioinformatics portal

شکل ۵-۷

بررسی یک رکورد نمونه

محتوای این صفحه (شکل ۵-۷) زیاد است. برای شرح جزییات این صفحه، نمونه ای کوچک تر را در صفحه بعد ارائه کرده ایم و در پایان این رکورد نیز توضیحات کاملی درباره آن داده ایم.

در نگاه اول، این اطلاعات بگرنج و نامفهوم به نظر می رسند. جزء جزء این اطلاعات را مورد بررسی قرار داده و مفهوم آن ها را شرح می دهیم. به این ساختار ارائه شده falt file گفته می شود. اطلاعات درون falt file ها به صورت Line type ارائه می شود. بدین معنا که در اول هر سطر، یک کلمه یا یک کد قرار گرفته است و بیان می کند در آن سطر چه اطلاعاتی گنجانده شده است. برای مثال در سطر اول کلمه LOCUS در ابتدای سطر، بیان می کند که در این سطر کد یکتای رکورد (SCU49845) یا Locus ID قرار دارد. یا در سطری که در ابتدای آن

کلمه ORGANIIM آمده، بیان کننده درخت تاکسانومی 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 Axl2p (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 Axl2p, 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

source Location/Qualifiers

1..5028

/organism="Saccharomyces cerevisiae"

/mol_type="genomic DNA"

/db_xref="taxon:[4932](#)"

/chromosome="IX"

/map="9"

[CDS](#) <1..206

/codon_start=3

/product="TCP1-beta"

/protein_id="[AAA98665.1](#)"

/db_xref="GI:1293614"

/translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLRRAVVSSASEAAEVLLRVDNIIRARPRTANRQHM"

[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"

/translation="MTQLQISLLL TATISLLHLVVATPYEAYPIGKQYPPVARVNESF
TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
VFNVTFDRSMFTNEESIVSYYGRS QLYNAPLPNWLFFDSGELKFTGTAPVINSIAIAPE
TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTD TGNVSYDLPLNYV
YLDDDPISSDKLGSINLLDAPDWVALDNATISGSPDELLGKNSNPANFSVSIYDTYG
DVIYFNFEVVSTTDLFAISSLPNINATRGWFSSYYFLPSQFTDYVNTNVSLEFTNSSQ
DHDWVKFQSSNLT LAGEVPKNFDKLSLGLKANQGSQS QELYFNIIGMDSKITHSNHSA
NATSTRSSHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
CGVAIPLGVILVALICFLIFWRRRRRENPD DENLPHAISGPD LNNPANKPNQENATPLN
NPFDDDASSYDDTSIARRLAALNTL KLDNHSATESDISSVDEKRDSLGMNTYNDQFQ
SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
YGSQKTVDTTEKLF DLEAPEKEKRTSRDVTMSSLDPWNSNISPSVRKSVTPSPYNVTK
HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
VDFSNSKNVNVGQVKDIHGRIPEML"

[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"

/translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
KDDQIITETEVFDEF RSSLNSLIMHLEKLPKVND DTTITFEAVINAIELELGHKLDRNR
RVDSLEEKAEIERDSNWWKCQEDENLPD NNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
LISGDDKILNGVYSQYEEGESIFGSLF"

ORIGIN

1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac
ggaaccattg
61 ccgacatgag acagttaggt atcgctcgaga gttacaagct aaaacgagca
gtagtcagct
121 ctgcatctga agccgctgaa gttctactaa ggggtggataa catcatccgt
gcaagaccaa
181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa
taataaaccg
241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta
tataattcaa

301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg
 tcacaaataa
 361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca
 agaattgtaat
 421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc
 tccttgccga
 481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat
 tttcttattc
 541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta
 gaagaacaga
 601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt
 ccaacatcta
 661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta
 ttgctgacag
 721 ctactatata actactccat ctagtagtgg ccacgcccta tgaggcatat
 cctatcgga
 781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt
 tccaatgata
 841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc
 gacttaccga
 901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct
 tctgacttac
 961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg
 gactctgccg
 1021 acagcacgtc tttgaacaat acataccaat ttgttggttac aaaccgtcca
 tccatctcgc
 1081 tatcgtcaga tttcaatcta ttggcggtgt taaaaaacta tgggtatact
 aacggcaaaa
 1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt
 tcaatgttca
 1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat
 gcgccgttac
 1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca
 ccggtgataa
 1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca
 gacattgaag
 1381 gattttctgc cgttgaggta gaattcgaat tagtcatcgg ggctcaccag
 ttaactacct
 1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca
 tatgacttac
 1501 ctctaaacta tgtttatctc gatgacgac ctatttcttc tgataaattg
 ggttctataa
 1561 acttattgga tgctccagac tgggtggcat tagataatgc taccatttcc
 gggctgtcc
 1621 cagatgaatt actcggtaag aactccaatc ctgccaattt ttctgtgtcc
 atttatgata
 1681 cttatggtga tgtgatttat 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 توالی های نوکلئوتیدی

Flat file ها به سه بخش اصلی تقسیم می شوند:

- (۱) حاوی اطلاعاتی (descriptors) است در مورد کل رکورد
- (۲) features (ویژگی ها) که یادداشت ها و توضیحات رکورد است.

(۳) و خود توالی نوکلئوتیدی.

همه flat file های پایگاه های 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 داده های نوکلئوتیدی مختلف تفاوت 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 نام های با معنا تاکنون استفاده شده است، نام های 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 نام ها 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 DNA یا RNA است.

• المان چهارم در این سطر (PLN)، کد تقسیم بندی (division code) است. این کد سه حرفی، کد تقسیم بندی تاکسونومی بوده یا برای دسته بندی های دیگری به کار می رفته است، NCBI کد تقسیم بندی تاکسونومی ارگانیزم ها 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 کدگذاری امروزه منسوخ شده است. جدیداً 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 توالی های کوتاه از cDNA هستند. معمولاً به تعداد زیادی از روی این نوع ژن رونویسی صورت می گیرد. EST ها نمایش یک تصویر فوری 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 توالی های کوتاه (بین ۲۰۰ الی ۵۰۰ جفت باز) و روی ژنوم یکتا هستند. جایگاه و توالی آن ها کاملاً مشخص است. این توالی ها قابل آشکار شدن به وسیله PCR هستند. توالی 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: 22-FEB-1998. اگر هر یک از ویژگی های رکورد به روز شده باشد و سپس رکورد انتشار یابد، این تاریخ، آخرین تاریخ انتشار است.

بخش بعدی header، سطر definition به فرمت زیر است:

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

در این سطر، اطلاعاتی ارائه 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 پایین تر رکورد که ببایم، 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 روبرو است: ۵+۱ و ۶+۲

"۵+۱" 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.
"۶+۲" 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 version 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 GI: 1293613 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 متأسفانه بی استفاده اند.

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

SOURCE *Saccharomyces cerevisiae* (baker's yeast)
ORGANISM [Saccharomyces cerevisiae](#)
Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.

هر رکورد GenBank باید حداقل یک مرجع (citation یا reference) داشته باشد. این مرجع ها مقالاتی هستند که به توالی اعتبار می دهند که چرا این ژن خاص تعیین توالی شده است. در بسیاری از موارد رکوردها دو یا چند مرجع دارند. نمونه ای از یک مرجع در زیر آمده است:

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](#)

اگر به رکورد اصلی مراجعه نمایید متوجه می شوید که سه مرجع با ساختار بالا وجود دارد. مرجع ها با شماره ای که دقیقاً جلوی کلمه REFERENCE قرار گرفته است از هم جدا شده اند. افزودن یک رکورد به بانک برای فرد اعتبار علمی ایجاد می کند. هر مرجعت از بخش های زیر تشکیل شده است.

• نویسندگان مقاله (AUTHORS)

• عنوان مقاله (TITLE)

• نام ژورنال و سال انتشار آن (JOURNAL)

• کد این مقاله (PMID) (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 مراجع ها به مقالات مرتبط با این رکورد دست یابید.

The Feature Table

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

FEATURES	Location/Qualifiers
source	1..5028 /organism="Saccharomyces cerevisiae" /mol_type="genomic DNA" /db_xref="taxon: 4932 " /chromosome="IX" /map="9"
CDS	<1..206 /codon_start=3 /product="TCP1-beta" /protein_id=" AAA98665.1 " /db_xref="GI:1293614"
	/translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA AEVLLRVDNIIRARPRTANRQHM"
gene	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 1...5028 نشان داده شده است.

- این ژن مربوط به " ساکارومایسس سروسیه " است که در سطر /organism در بخش source 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 chromosome/ در زیر source 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 يك CDS جداگانه گذاشته شده است. همچنان که در قسمت اولین CDS می بینید اولین اگزون از باز ۱ شروع و به باز ۲۰۶ ختم می شود. بدون شروع آن باز شماره ۳ است (/codon_start=3). نام این پروتئین TCP1-beta، شناسه آن AAA98665 در سطرهای بعدی آمده است. ترجمه این exon نیز در سطر translation مشاهده می نمایید.
- نام این ژن AXL2، در سطر gene قرار دارد.

توالی ژن ها

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 داده ایم. در بسیاری از نرم افزارهای کامپیوتری از این توالی نوکلئوتیدی به عنوان ورودی نرم افزار استفاده می شود مانند نرم افزار BLAST که توالی های نوکلئوتیدی را با یکدیگر مقایسه می نماید.

نکته آخر در این فصل :

احتمالا 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 مثلا توی عکسهای کامپیوتر و یا موزیک های کامپیوتری با فرمتهای زیادی آشنا هستید

در دادههای بیوانفورماتیکی فرمتهای خاصی برای ذخیره کردن اطلاعات وجود دارد .

در بالای صفحه ای که توالی ژن را بررسی کردیم اگر توجه کرده باشید بخش Display را دیده اید که برای تغییر

دادهها به فرمتهای مختلف میباشد

NCBI Nucleotide

Search CoreNucleotide for [] Go Clear

Limits Preview/Index History Clipboard Detail

Display GenBank Show 5 Send to Hide: ☐ sequence ☐ all but gene, CDS and mRNA features

Range: Brief Summary GI List ASN.1 GenBank GenBank(Full) **FASTA** XML TinySeq XML INSDSeq XML Graph

1: U49845.1 Saccharomyces cer...[gi:1293613]

LOCUS linear PLN 21-JUN-1999

DEFINITION: Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p Rev7p (REV7) genes, complete cds.

ACCESSION: U49845.1

VERSION: U49845.1 GI:1293613

KEYWORDS:

SOURCE: Saccharomyces cerevisiae (baker's yeast)

WWW.IBP.IR iranain bioinformatics portal

شکل ۶-۷

NCBI Nucleotide

Search CoreNucleotide for [] Go Clear

Limits Preview/Index History Clipboard Details

Display **FASTA** Show 5 Send to

Range: begin to end ☐ Reverse complemented strand Refresh

U49845.1 Reports Saccharomyces cer...[gi:1293613]

U49845.1|SCU49845 Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds

Links

مشاهده میکنید که با انتخاب فرمت FASTA اطلاعات و توضیحات بالای توالی حذف شده و توالی ژنوم بدون دسته بندی ده تایی ردیف شده اند که حال میتوانید از اطلاعات به این فرمت در نرم افزارهای بیوانفورماتیکی استفاده کنید

WWW.IBP.IR iranain bioinformatics portal

شکل ۷-۷

منابع:

۱. ابزارهای بیوانفورماتیک نویسندگان: فاطمه هنری و حبیب اسلام زاده
برای اطلاعات بیشتر در رابطه با این کتاب میتوانید به آدرس 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 نوشته های 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**

دوست ارجمند این مطلب آموزشی به منظور آشنایی هر چه بیشتر ایرانیان با علم بیوانفورماتیک
تهیه شده است خواهشمند است از هر گونه سوء استفاده از این مطالب پرهیزید زیرا که این
مطالب تحت حمایت قانون کپی رایت میباشد

این مطالب به این دلیل به صورت کتابهای الکترونیکی تهیه شده اند تا
هرکس، در هر جا، در هر زمان و به صورت رایگان
بتواند به این مطالب دست رسی داشته باشد پس اگر چنانچه تمایل به استفاده از این مطالب در
سایت یا وبلاگ خود را دارید این کار با ذکر منبع بلامانع میباشد.

هیچ شخص حقیقی یا حقوقی حق ندارد تا از مطالب کتب الکترونیکی این پرتال به منظور چاپ
و نشر کتاب استفاده کند چاپ و نشر کتاب کاری مقدس میباشد ولی فروش علم جزء ناشایسته
ترین کارهاست که متأسفانه برخی با استفاده از نا آشنایی مردم با منبع سرشار اینترنت دست به
چاپ کتب میزنند و آن را به قیمت بسیار بالایی به فروش میرسانند

صحبت این است که اگر عاشق علم هستید دانش خود را به صورت رایگان در اختیار همگان
قرار دهید و بهترین راه آن توسط اینترنت میباشد.

به امید روزی که هیچ کتابی چاپ نشود و تمام مطالب به صورت رایگان در اینترنت قابل دست
رسی باشند تا حد اقل در بخش استفاده از منابع علمی نامی از غنی و فقیر برده نشود .

پرتال بیوانفورماتیک ایرانیان www.ibp.ir

بابک باباعباسی

WWW.IBP.IR

iranian bioinformatics portal