BIOINFORMATICS

Course: Molecular Biology (02022312)

Instructor: Dr. M A Srour

Textbook:
Watson J, Baker TA, Bell SP, Gann A, Levine M, Losick R (2008). Molecular Biology of the Gene, 6th ed. , Chap 20 pp. 703-31; chap 21 pp.739-82.
Malacinski GM (2003). Essentials of Molecular Biology, 4th ed. Chap 13, pp. 284-301

Lec # 14 Wed 25.04.2012

Bioinformatics (Computational Biology)

- Bioinformatics: study of the meaning (interpretation) of information contained in an organism's genome
- Or bioinformatics is now routinely applied to those molecular biology endeavors that rely heavily on highly sophisticated computations
- Ability to rapidly sequence DNA has led to large databases
  - Development of new algorithms (formulae for solving complex, highly specific problems)
  - Data analysis and interpretation

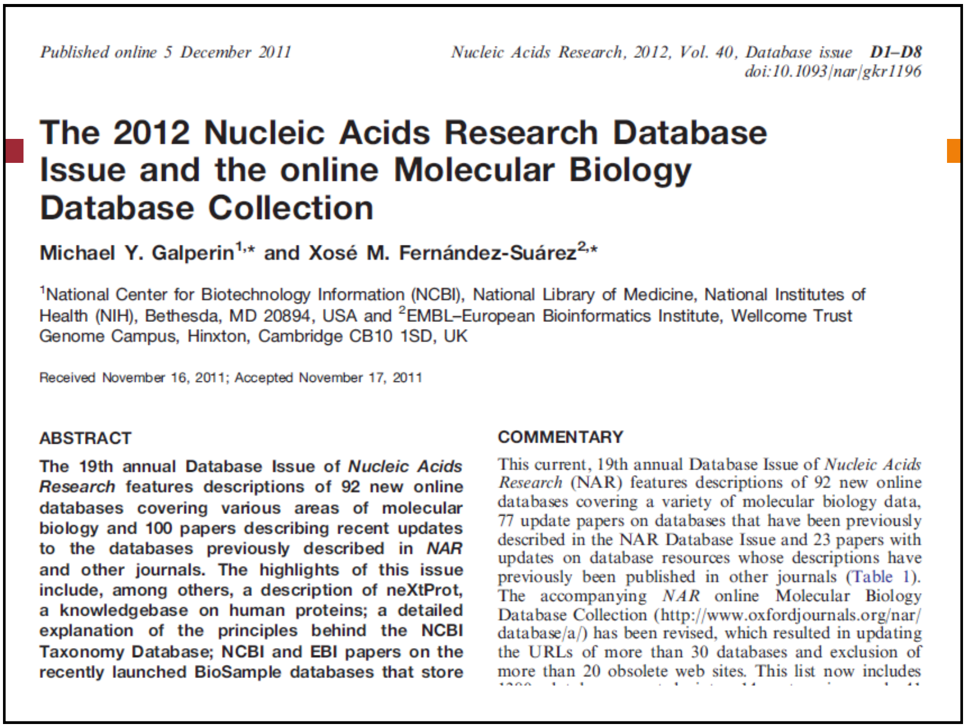
Examples of bioinformatics programs

(Malacinski 2003/Tables 13-1 & 13-2)

Program	Example of Type of research task
Genomics	
Gene counting	Identify nucleotide sequences that serve protein-coding function
	Estimate gene numbers to understand levels of complexity of regulatory mechanisms
Medical genomics	Compare disease-causing genes in model organisms (e.g., mice) with their human counterparts to identify structural homologies at both DNA and protein level
Proteomics	
Identification of sample protein	Use amino acid compositions of newly discovered protein to compare with known proteins
	Use the nucleotide sequence to compare with known genes
	Translate a nucleotide sequence into amino acid sequence

Molecular Biology Databases

- Database: it's a repository of information that has a specific structure that enables the entering and extraction of data and in many cases also aids analysis of the data
- The list of Molecular Biology Databases (URL) is continuously expanding, and updates are published regularly by the journal of "Nucleic Acid Research" (1380 databases are now listed in 2012: at the web site of NAR online Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/a/)



Published online 5 December 2011

Nucleic Acids Research, 2012, Vol. 40, Database issue ***D1–D8***
doi:10.1093/nar/gkr1196

The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection

Michael Y. Galperin[1,*] **and Xosé M. Fernández-Suárez**[2,*]

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA and [2]EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received November 16, 2011; Accepted November 17, 2011

ABSTRACT

The 19th annual Database Issue of *Nucleic Acids Research* features descriptions of 92 new online databases covering various areas of molecular biology and 100 papers describing recent updates to the databases previously described in *NAR* and other journals. The highlights of this issue include, among others, a description of neXtProt, a knowledgebase on human proteins; a detailed explanation of the principles behind the NCBI Taxonomy Database; NCBI and EBI papers on the recently launched BioSample databases that store

COMMENTARY

This current, 19th annual Database Issue of *Nucleic Acids Research* (NAR) features descriptions of 92 new online databases covering a variety of molecular biology data, 77 update papers on databases that have been previously described in the NAR Database Issue and 23 papers with updates on database resources whose descriptions have previously been published in other journals (Table 1). The accompanying *NAR* online Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/a/) has been revised, which resulted in updating the URLs of more than 30 databases and exclusion of more than 20 obsolete web sites. This list now includes

Molecular Biology Databases

- Three primary Nucleotide Sequence Databases
  - GenBank (http://www.ncbi.nlm.nih.gov/genbank), National Center for Biotechnology Information, USA
  - EMBL nucleotide sequence database (http://www.ebi.ac.uk/embl/), European Molecular Biology Laboratory
  - DNA Data Bank of Japan, DDBJ (http://www.ddbj.nig.ac.jp/)
  - Many other resources like RNA-specific databases,..

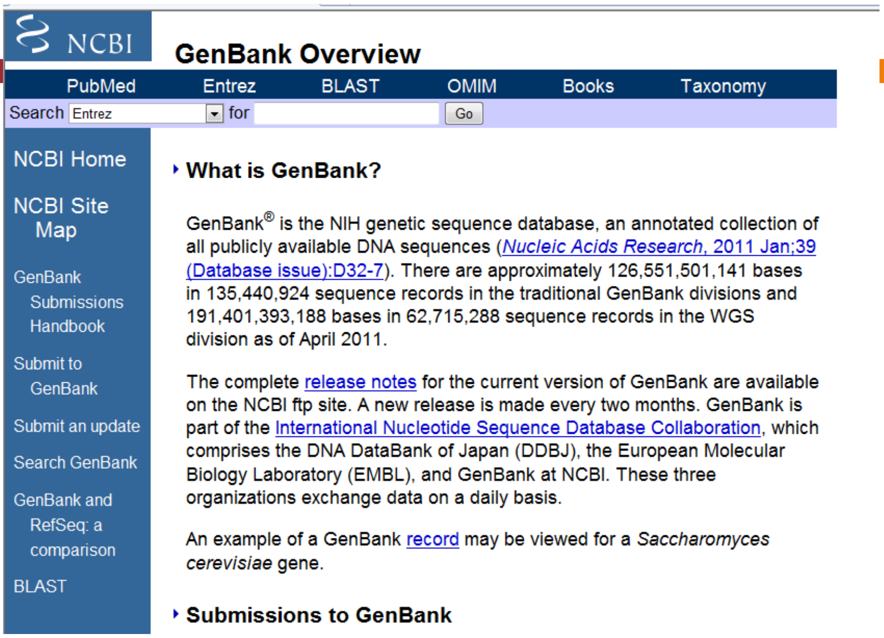
Molecular Biology Databases

- Types of DNA sequences stored in these databases
  - Raw genomic sequence representing chromosomal DNA
  - cDNAs , synthesized by reverse transcription of mRNA molecules
  - Expressed sequence tags (ESTs), partial cDNA sequences

Molecular Biology Databases

- Protein sequence databases, such as
  - UniProtKB & its annotated component Swiss-Prot, produced by analysis of all translations of EMBL
  - NCBI protein database
- Annotated to include ancillary information (author, publications, etc.)

☐ Annotation (sequence & genome): the systematic identification of every stretch of genomic DNA that contains protein-coding information or noncoding sequence that specify regulatory RNAs such as microRNAs (miRNAs) http://www.ncbi.nlm.nih.gov/genbank/



GenBank Overview

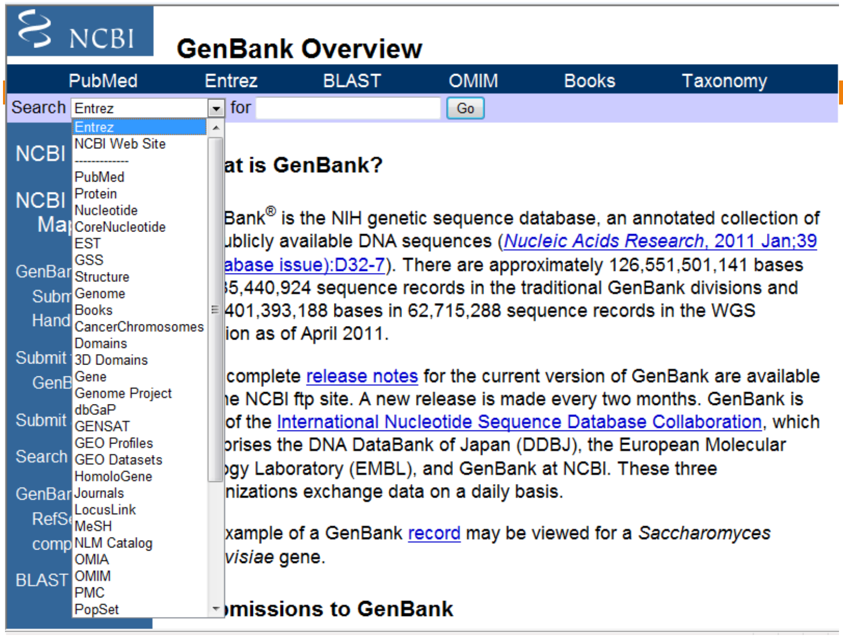
What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2011 Jan;39 (Database issue):D32-7). There are approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the WGS division as of April 2011.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank record may be viewed for a *Saccharomyces cerevisiae* gene.

Submissions to GenBank



GenBank Overview

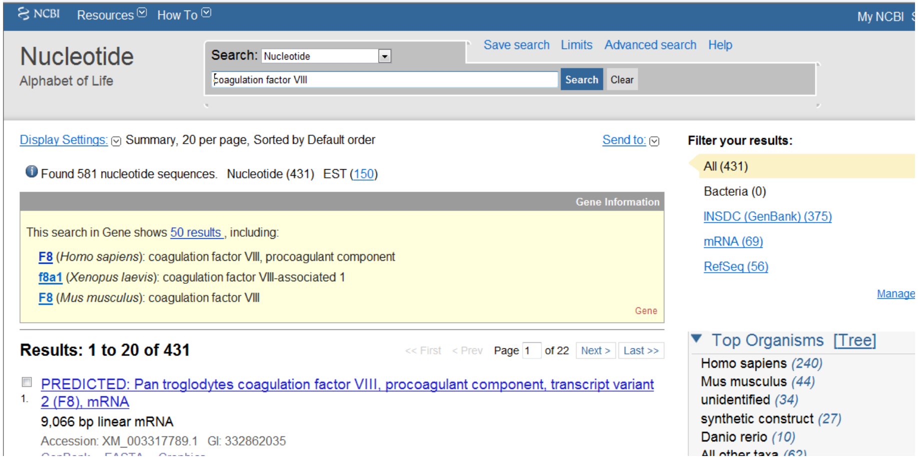
at is GenBank?

Bank® is the NIH genetic sequence database, an annotated collection of
ublicly available DNA sequences (*Nucleic Acids Research*, 2011 Jan;39
abase issue):D32-7). There are approximately 126,551,501,141 bases
5,440,924 sequence records in the traditional GenBank divisions and
401,393,188 bases in 62,715,288 sequence records in the WGS
ion as of April 2011.

complete release notes for the current version of GenBank are available
e NCBI ftp site. A new release is made every two months. GenBank is
of the International Nucleotide Sequence Database Collaboration, which
prises the DNA DataBank of Japan (DDBJ), the European Molecular
ogy Laboratory (EMBL), and GenBank at NCBI. These three
nizations exchange data on a daily basis.

xample of a GenBank record may be viewed for a *Saccharomyces*
visiae gene.

missions to GenBank



Nucleotide
Alphabet of Life

Search: Nucleotide

coagulation factor VIII

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

Send to:

Found 581 nucleotide sequences. Nucleotide (431) EST (150)

Gene Information

This search in Gene shows 50 results , including:

F8 (*Homo sapiens*): coagulation factor VIII, procoagulant component
f8a1 (*Xenopus laevis*): coagulation factor VIII-associated 1
F8 (*Mus musculus*): coagulation factor VIII

Gene

Results: 1 to 20 of 431

<< First < Prev Page 1 of 22 Next > Last >>

1. PREDICTED: Pan troglodytes coagulation factor VIII, procoagulant component, transcript variant 2 (F8), mRNA
9,066 bp linear mRNA
Accession: XM_003317789.1 GI: 332862035

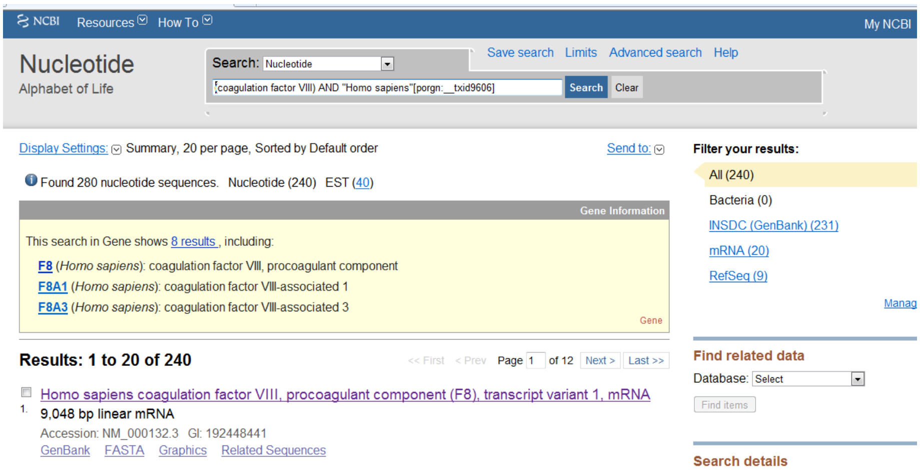
Filter your results:

All (431)
Bacteria (0)
INSDC (GenBank) (375)
mRNA (69)
RefSeq (56)

Manage

Top Organisms [Tree]

Homo sapiens (240)
Mus musculus (44)
unidentified (34)
synthetic construct (27)
Danio rerio (10)
All other taxa (62)



Nucleotide
Alphabet of Life

Search: Nucleotide

(coagulation factor VIII) AND "Homo sapiens"[porgn:__txid9606]

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

Send to:

Found 280 nucleotide sequences. Nucleotide (240) EST (40)

Gene Information

This search in Gene shows 8 results , including:

F8 (*Homo sapiens*): coagulation factor VIII, procoagulant component
F8A1 (*Homo sapiens*): coagulation factor VIII-associated 1
F8A3 (*Homo sapiens*): coagulation factor VIII-associated 3

Gene

Results: 1 to 20 of 240

<< First < Prev Page 1 of 12 Next > Last >>

1. Homo sapiens coagulation factor VIII, procoagulant component (F8), transcript variant 1, mRNA
9,048 bp linear mRNA
Accession: NM_000132.3 GI: 192448441
GenBank FASTA Graphics Related Sequences

Filter your results:

All (240)
Bacteria (0)
INSDC (GenBank) (231)
mRNA (20)
RefSeq (9)

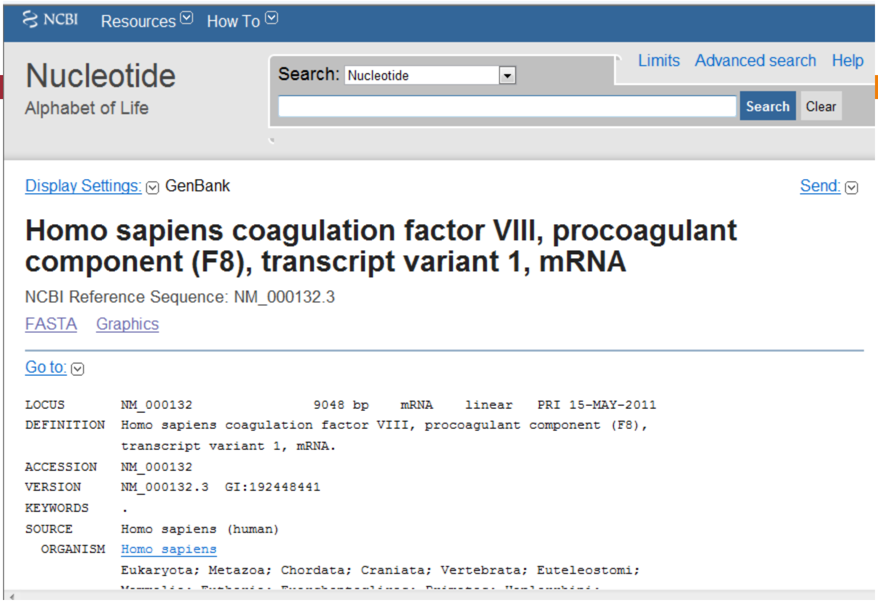
Manag

Find related data

Database: Select

Find items

Search details

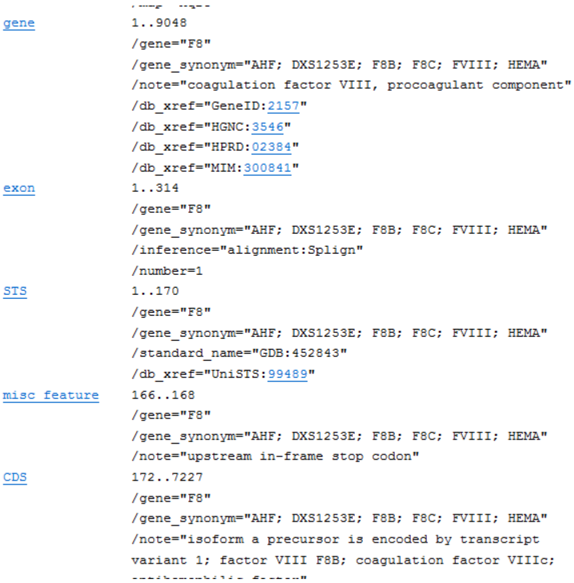


Display Settings: GenBank

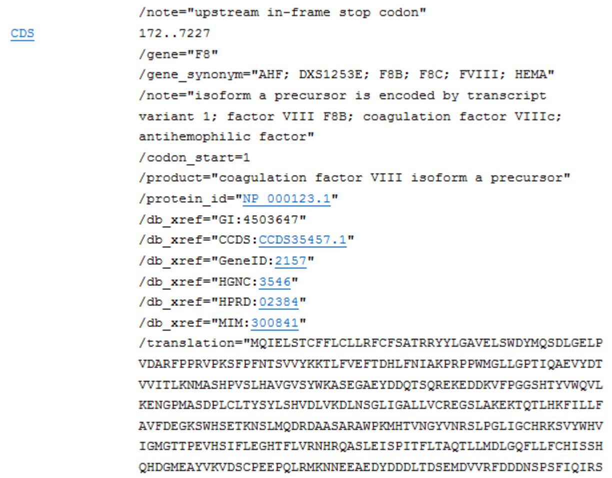
Homo sapiens coagulation factor VIII, procoagulant component (F8), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000132.3

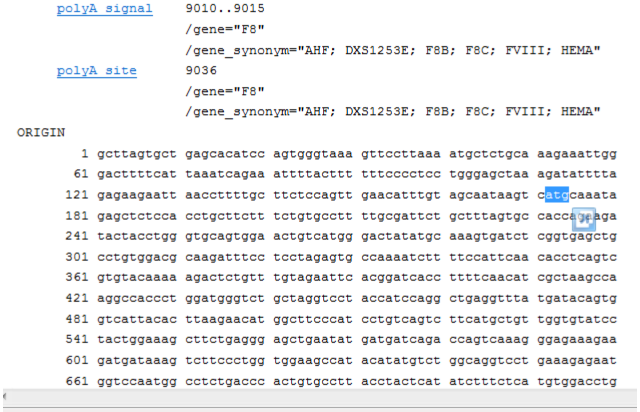
```
LOCUS       NM_000132               9048 bp    mRNA    linear   PRI 15-MAY-2011
DEFINITION  Homo sapiens coagulation factor VIII, procoagulant component (F8),
            transcript variant 1, mRNA.
ACCESSION   NM_000132
VERSION     NM_000132.3  GI:192448441
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
```



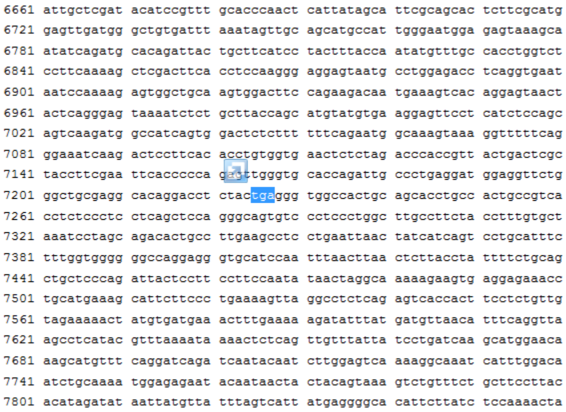
```
     gene            1..9048
                     /gene="F8"
                     /gene_synonym="AHF; DXS1253E; F8B; F8C; FVIII; HEMA"
                     /note="coagulation factor VIII, procoagulant component"
                     /db_xref="GeneID:2157"
                     /db_xref="HGNC:3546"
                     /db_xref="HPRD:02384"
                     /db_xref="MIM:300841"
     exon            1..314
                     /gene="F8"
                     /gene_synonym="AHF; DXS1253E; F8B; F8C; FVIII; HEMA"
                     /inference="alignment:Splign"
                     /number=1
     STS             1..170
                     /gene="F8"
                     /gene_synonym="AHF; DXS1253E; F8B; F8C; FVIII; HEMA"
                     /standard_name="GDB:452843"
                     /db_xref="UniSTS:99489"
     misc_feature    166..168
                     /gene="F8"
                     /gene_synonym="AHF; DXS1253E; F8B; F8C; FVIII; HEMA"
                     /note="upstream in-frame stop codon"
     CDS             172..7227
                     /gene="F8"
                     /gene_synonym="AHF; DXS1253E; F8B; F8C; FVIII; HEMA"
                     /note="isoform a precursor is encoded by transcript
                     variant 1; factor VIII F8B; coagulation factor VIIIc;
```



```
                     /note="upstream in-frame stop codon"
     CDS             172..7227
                     /gene="F8"
                     /gene_synonym="AHF; DXS1253E; F8B; F8C; FVIII; HEMA"
                     /note="isoform a precursor is encoded by transcript
                     variant 1; factor VIII F8B; coagulation factor VIIIc;
                     antihemophilic factor"
                     /codon_start=1
                     /product="coagulation factor VIII isoform a precursor"
                     /protein_id="NP_000123.1"
                     /db_xref="GI:4503647"
                     /db_xref="CCDS:CCDS35457.1"
                     /db_xref="GeneID:2157"
                     /db_xref="HGNC:3546"
                     /db_xref="HPRD:02384"
                     /db_xref="MIM:300841"
                     /translation="MQIELSTCFFLCLLRFCFSATRRYYLGAVELSWDYMQSDLGELP
                     VDARFPPRVPKSFPFNTSVVYKKTLFVEFTDHLFNIAKPRPPWMGLLGPTIQAEVYDT
                     VVITLKNMASHPVSLHAVGVSYWKASEGAEYDDQTSQREKEDDKVFPGGSHTYVWQVL
                     KENGPMASDPLCLTYSYLSHVDLVKDLNSGLIGALLVCREGSLAKEKTQTLHKFILLF
                     AVFDEGKSWHSETKNSLMQDRDAASARAWPKMHTVNGYVNRSLPGLIGCHRKSVYWHV
                     IGMGTTPEVHSIFLEGHTFLVRNHRQASLEISPITFLTAQTLLMDLGQFLLFCHISSH
                     QHDGMEAYVKVDSCPEEPQLRMKNNEEAEDYDDDLTDSEMDVVRFDDDNSPSFIQIRS
```



```
     polyA_signal    9010..9015
                     /gene="F8"
                     /gene_synonym="AHF; DXS1253E; F8B; F8C; FVIII; HEMA"
     polyA_site      9036
                     /gene="F8"
                     /gene_synonym="AHF; DXS1253E; F8B; F8C; FVIII; HEMA"
ORIGIN
        1 gcttagtgct gagcacatcc agtgggtaaa gttccttaaa atgctctgca aagaaattgg
       61 gacttttcat taaatcagaa attttacttt tttcccctcc tgggagctaa agatatttta
      121 gagaagaatt aacctttgc ttctccagtt gaacatttgt agcaataagt catgcaaata
      181 gagctctcca cctgcttctt tctgtgcctt ttgcgattct gctttagtgc caccagaaga
      241 tactacctgg gtgcagtgga actgtcatgg gactatatgc aaagtgatct cggtgagctg
      301 cctgtggacg caagatttcc tcctagagtg ccaaaatctt ttccattcaa cacctcagtc
      361 gtgtacaaaa agactctgtt tgtagaattc acggatcacc ttttcaacat cgctaagcca
      421 aggccaccct ggatgggtct gctaggtcct accatccagg ctgaggttta tgatacagtg
      481 gtcattacac ttaagaacat ggcttcccat cctgtcagtc ttcatgctgt tggtgtatcc
      541 tactggaaag cttctgaggg agctgaatat gatgatcaga ccagtcaaag ggagaaagaa
      601 gatgataaag tcttccctgg tggaagccat acatatgtct ggcaggtcct gaaagagaat
      661 ggtccaatgg cctctgaccc actgtgcctt acctactcat atctttctca tgtggacctg
```



```
6661 attgctcgat acatccgttt gcacccaact cattatagca ttcgcagcac tcttcgcatg
6721 gagttgatgg gctgtgattt aaatagttgc agcatgccat tgggaatgga gagtaaagca
6781 atatcagatg cacagattac tgcttcatcc tactttacca atatgtttgc cacctggtct
6841 ccttcaaaag ctcgacttca cctccaaggg aggagtaatg cctggagacc tcaggtgaat
6901 aatccaaaag agtggctgca agtggacttc cagaagacaa tgaaagtcac aggagtaact
6961 actcagggag taaaatctct gcttaccagc atgtatgtga aggagttcct catctccagc
7021 agtcaagatg gccatcagtg gactctcttt tttcagaatg gcaaagtaaa ggtttttcag
7081 ggaaatcaag actccttcac acctgtggtg aactctctag acccaccgtt actgactcgc
7141 taccttcgaa ttcacccca gacttgggtg caccagattg ccctgaggat ggaggttctg
7201 ggctgcgagg cacaggacct ctactgaggg tggccactgc agcacctgcc actgccgtca
7261 cctctccctc ctcagctcca gggcagtgtc cctccctggc ttgccttcta cctttgtgct
7321 aaatcctagc agacactgcc ttgaagcctc ctgaattaac tatcatcagt cctgcatttc
7381 tttggtgggg ggccaggagg gtgcatccaa tttaacttaa ctcttaccta ttttctgcag
7441 ctgctcccag attactcctt ccttccaata taactaggca aaaagaagtg aggagaaacc
7501 tgcatgaaag cattcttccc tgaaaagtta ggcctctcag agtcaccact tcctctgttg
7561 tagaaaaact atgtgatgaa actttgaaaa agatatttat gatgttaaca tttcaggtta
7621 agcctcatac gtttaaaata aaactctcag ttgtttatta tcctgatcaa gcatggaaca
7681 aagcatgttt caggatcaga tcaatacaat cttggagtca aaaggcaaat catttggaca
7741 atctgcaaaa tggagagaat acaataacta ctacagtaaa gtctgtttct gcttccttac
7801 acatagatat aattatgtta tttagtcatt atgaggggca cattcttatc tccaaaacta
```

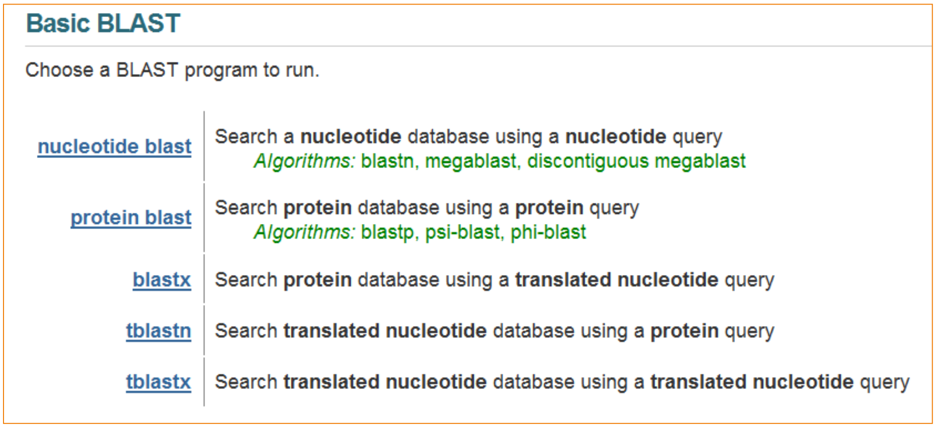
Searching databases & principles of sequence alignment

- Alignment: the task of locating equivalent regions of two or more sequences to maximize their similarity
- Fundamental question in alignment: is the similarities observed due to "chance" and thus of little biological significance or due to the derivation from common ancestral sequence (or homologous)
- Homology or similarity reflect a common ancestor = common function or structure
- Homology: refers to similarity in sequence or structure due to descent from a common ancestor

Searching databases & principles of sequence alignment

- Searching databases needs to be sensitive (avoid false negative) and specific (avoid false positive)
- Data-base-search algorithm: BLAST (Basic local Alignment Search Tool)
- BLAST is one of the most widely used search programs
- BLAST relies on finding core similarity, which is defined by a window of preset size (called a "word") with a certain minimum density of matches (for DNA) or with an amino acid similarity score above threshold (for proteins)

Searching databases:
Basic Local Alignment Search Tool (BLAST)



Basic BLAST

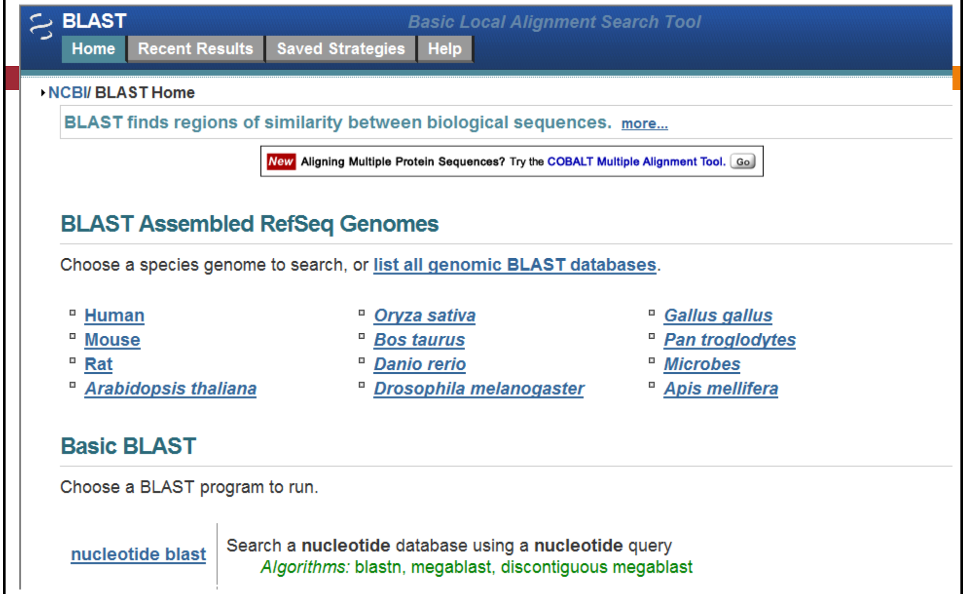
Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Source: http://blast.ncbi.nlm.nih.gov/Blast.cgi

Doing a BLAST Search

- http://www.ncbi.nlm.nih.gov/ BLAST/
- choose BLAST program
- paste in query sequence or acc. no. → BLAST!
- change default options:
  - database (nr = non-redundant)
  - scoring matrix and gap penalties
  - filtering
  - E-value cutoff (ie, Expectation value): the smaller the E-value, the better the alignment
  - limit subset of database (organism, keyword, etc.)
  - display options (eg, # of descriptions, alignments, etc.)



BLAST finds regions of similarity between biological sequences. more...

Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool.

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or list all genomic BLAST databases.

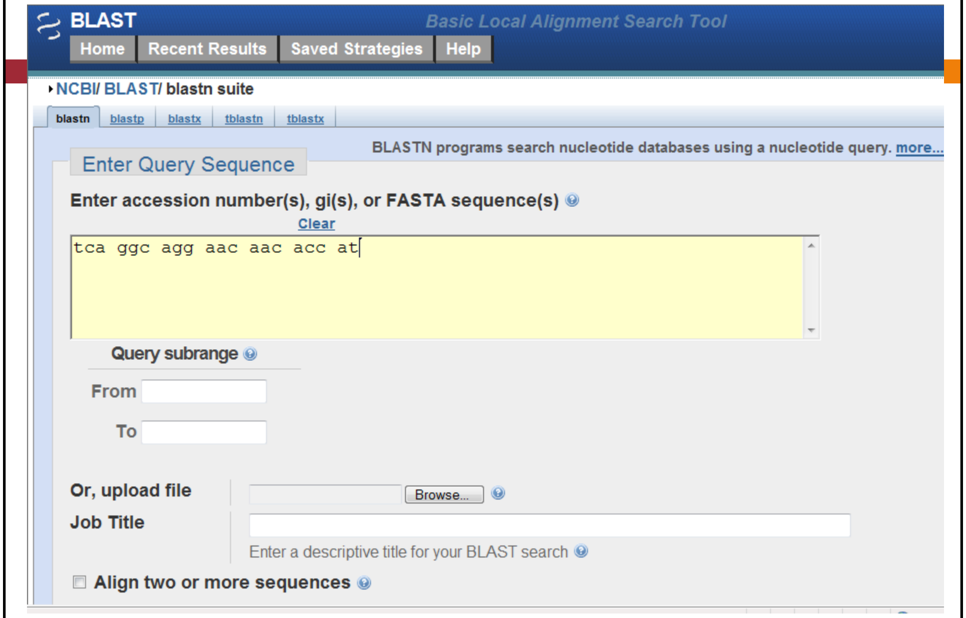
- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast

Search fro FV sequence (shown here is the sense primer)



BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

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

Clear

```
tca ggc agg aac aac acc at
```

Query subrange

From

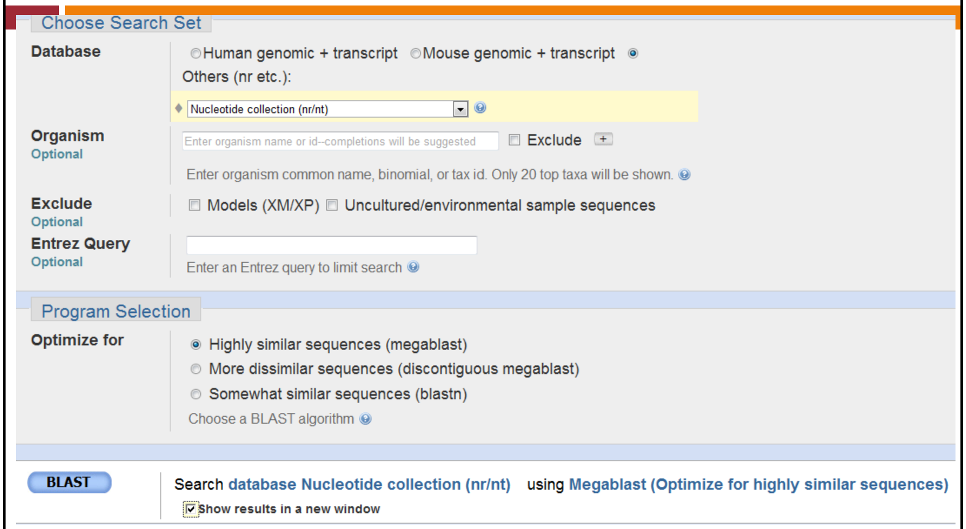
To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences



Choose Search Set

Database ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional

Enter organism name or id--completions will be suggested ☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

● Highly similar sequences (megablast)

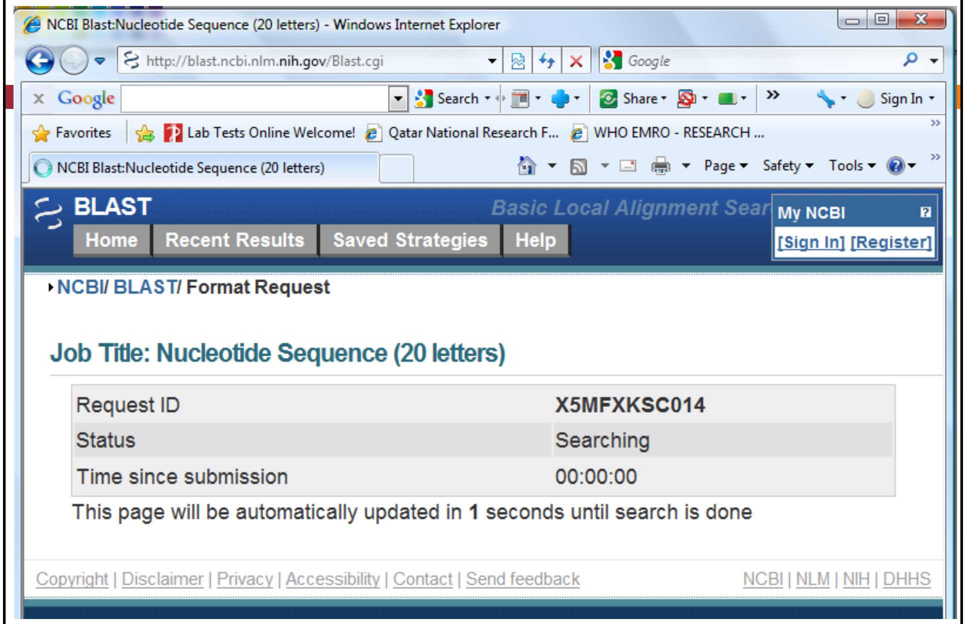
○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

☑ Show results in a new window

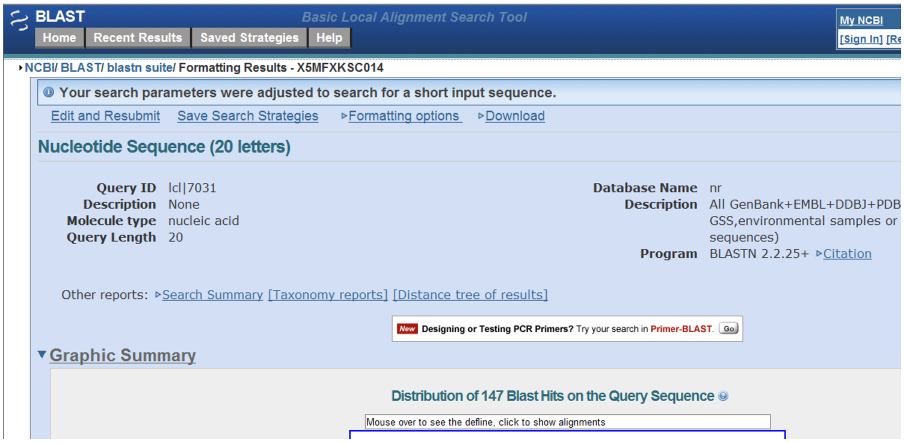


NCBI/ BLAST/ **Format Request**

Job Title: Nucleotide Sequence (20 letters)

Request ID	**X5MFXKSC014**
Status	Searching
Time since submission	00:00:00

This page will be automatically updated in **1** seconds until search is done



NCBI/ BLAST/ blastn suite/ **Formatting Results - X5MFXKSC014**

Your search parameters were adjusted to search for a short input sequence.

Nucleotide Sequence (20 letters)

Query ID lcl|7031
Description None
Molecule type nucleic acid
Query Length 20

Database Name nr
Description All GenBank+EMBL+DDBJ+PDB GSS,environmental samples or sequences)
Program BLASTN 2.2.25+ Citation

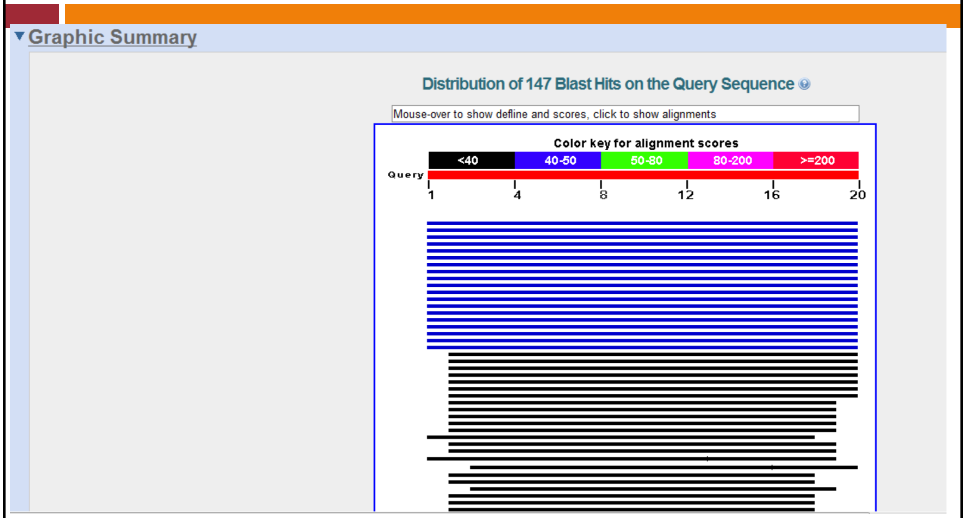
Other reports: Search Summary [Taxonomy reports] [Distance tree of results]

New **Designing or Testing PCR Primers?** Try your search in Primer-BLAST. Go

Graphic Summary

Distribution of 147 Blast Hits on the Query Sequence

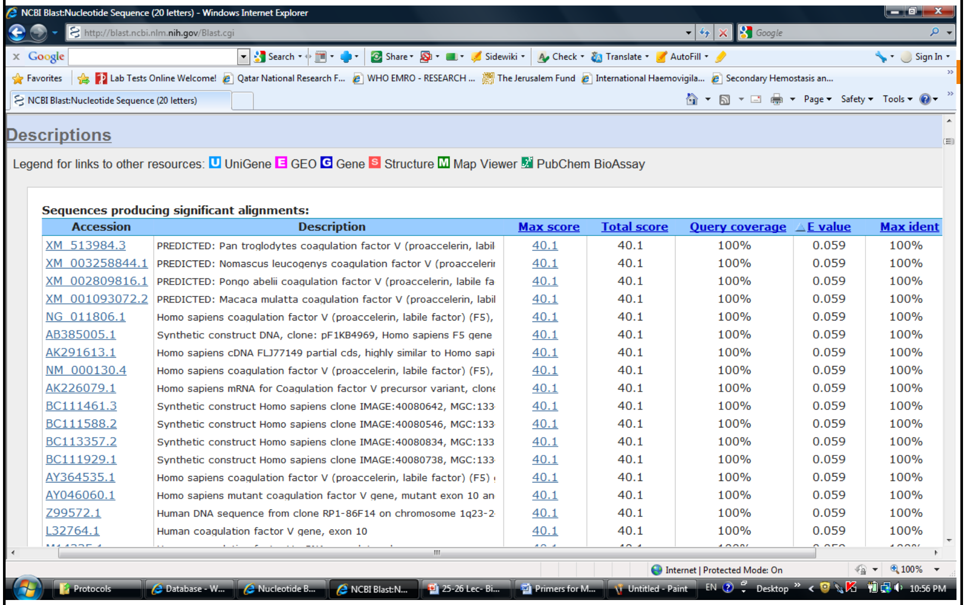
Mouse over to see the define, click to show alignments



Graphic Summary

Distribution of 147 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

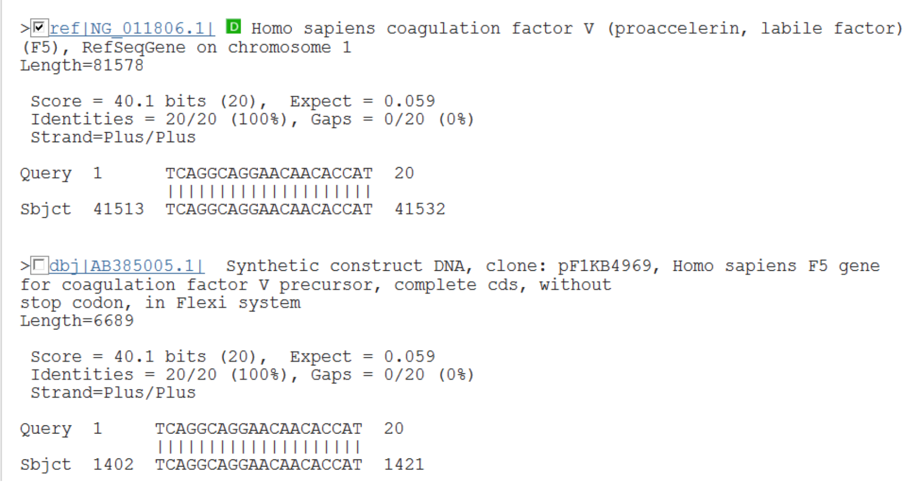


Descriptions

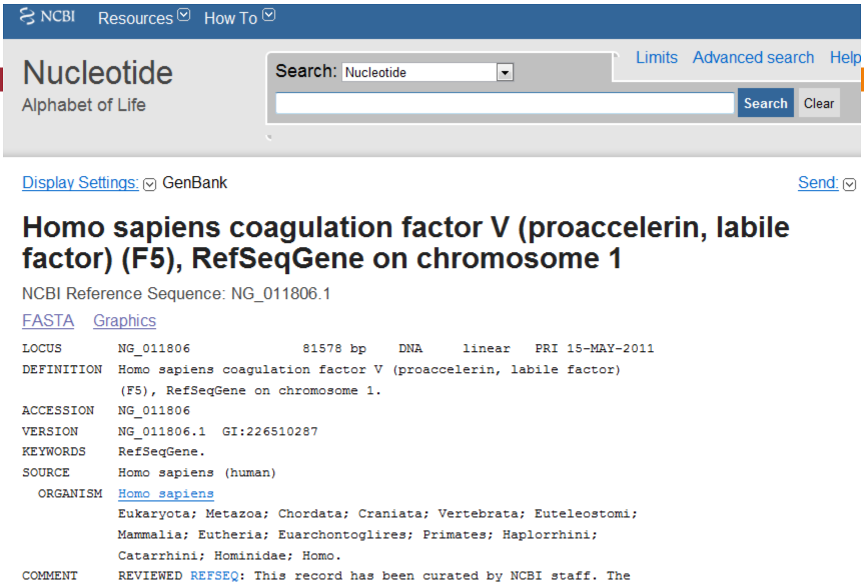
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
XM_513984.3	PREDICTED: Pan troglodytes coagulation factor V (proaccelerin, labil	40.1	40.1	100%	0.059	100%
XM_003258844.1	PREDICTED: Nomascus leucogenys coagulation factor V (proacceleri	40.1	40.1	100%	0.059	100%
XM_002809816.1	PREDICTED: Pongo abelii coagulation factor V (proaccelerin, labile fa	40.1	40.1	100%	0.059	100%
XM_001093072.2	PREDICTED: Macaca mulatta coagulation factor V (proaccelerin, labil	40.1	40.1	100%	0.059	100%
NG_011806.1	Homo sapiens coagulation factor V (proaccelerin, labile factor) (F5),	40.1	40.1	100%	0.059	100%
AB385005.1	Synthetic construct DNA, clone: pF1KB4969, Homo sapiens F5 gene	40.1	40.1	100%	0.059	100%
AK291613.1	Homo sapiens cDNA FLJ77149 partial cds, highly similar to Homo sapi	40.1	40.1	100%	0.059	100%
NM_000130.4	Homo sapiens coagulation factor V (proaccelerin, labile factor) (F5),	40.1	40.1	100%	0.059	100%
AK226079.1	Homo sapiens mRNA for Coagulation factor V precursor variant, clone	40.1	40.1	100%	0.059	100%
BC111461.3	Synthetic construct Homo sapiens clone IMAGE:40080642, MGC:133	40.1	40.1	100%	0.059	100%
BC111588.2	Synthetic construct Homo sapiens clone IMAGE:40080546, MGC:133	40.1	40.1	100%	0.059	100%
BC113357.2	Synthetic construct Homo sapiens clone IMAGE:40080834, MGC:133	40.1	40.1	100%	0.059	100%
BC111929.1	Synthetic construct Homo sapiens clone IMAGE:40080738, MGC:133	40.1	40.1	100%	0.059	100%
AY364535.1	Homo sapiens coagulation factor V (proaccelerin, labile factor) (F5)	40.1	40.1	100%	0.059	100%
AY046060.1	Homo sapiens mutant coagulation factor V gene, mutant exon 10 an	40.1	40.1	100%	0.059	100%
Z99572.1	Human DNA sequence from clone RP1-86F14 on chromosome 1q23-2	40.1	40.1	100%	0.059	100%
L32764.1	Human coagulation factor V gene, exon 10	40.1	40.1	100%	0.059	100%



```
>☑ref|NG_011806.1| D Homo sapiens coagulation factor V (proaccelerin, labile factor)
(F5), RefSeqGene on chromosome 1
Length=81578

 Score = 40.1 bits (20),  Expect = 0.059
 Identities = 20/20 (100%), Gaps = 0/20 (0%)
 Strand=Plus/Plus

Query  1      TCAGGCAGGAACAACACCAT  20
              ||||||||||||||||||||
Sbjct  41513  TCAGGCAGGAACAACACCAT  41532

>☐dbj|AB385005.1|  Synthetic construct DNA, clone: pF1KB4969, Homo sapiens F5 gene
for coagulation factor V precursor, complete cds, without
stop codon, in Flexi system
Length=6689

 Score = 40.1 bits (20),  Expect = 0.059
 Identities = 20/20 (100%), Gaps = 0/20 (0%)
 Strand=Plus/Plus

Query  1     TCAGGCAGGAACAACACCAT  20
             ||||||||||||||||||||
Sbjct  1402  TCAGGCAGGAACAACACCAT  1421
```



Nucleotide

Alphabet of Life

Search: Nucleotide

Display Settings: GenBank

Homo sapiens coagulation factor V (proaccelerin, labile factor) (F5), RefSeqGene on chromosome 1

NCBI Reference Sequence: NG_011806.1

```
LOCUS       NG_011806              81578 bp    DNA     linear   PRI 15-MAY-2011
DEFINITION  Homo sapiens coagulation factor V (proaccelerin, labile factor)
            (F5), RefSeqGene on chromosome 1.
ACCESSION   NG_011806
VERSION     NG_011806.1  GI:226510287
KEYWORDS    RefSeqGene.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
```

Blast Search Results

Probability

```
Query= Pbpp58b (423 letters)
Database: nr (493,611 sequences; 154,780,071 total letters)
                                                                Score     E
Sequences producing significant alignments:                     (bits)  Value

sp|Q08168|HRP_PLABE 58 KD PHOSPHOPROTEIN (HEAT SHOCK-RELATED PRO...   334   1e-90
gb|AAC37300.1| (L21710) 58 kDa phosphoprotein [Plasmodium berghei]   329   3e-89
pir||T10455 heat shock related protein - Plasmodium berghei >gi|...   250   2e-65
sp|P50503|HIP_RAT HSC70-INTERACTING PROTEIN >gi|4379408|emb|CAA5...   106   5e-22
sp|P50502|HIP_HUMAN HSC70-INTERACTING PROTEIN (PROGESTERONE RECE...    87   3e-16
gb|AAF45894.1| (AE003429) CG2947 gene product [Drosophila melano...    87   4e-16
pir||T24865 hypothetical protein T12D8.8 - Caenorhabditis elegan...    86   5e-16
pir||T04562 hypothetical protein T12H17.60 - Arabidopsis thalian...    81   2e-14
.
.
.
.
.
emb|CAA61595.1| (X89416) protein phosphatase 5 [Homo sapiens]         43   0.007
pdb|1A17|    Tetratricopeptide Repeats Of Protein Phosphatase 5       43   0.007
ref|NP_006238.1|| protein phosphatase 5, catalytic subunit >gi|1...   43   0.007
pir||S52570 phosphoprotein phosphatase (EC 3.1.3.16) 5, catalyti...   43   0.007
```

database | accession # | entry name or locus

Comparison of human NF1 and yeast *S. cerevisiea* Ira protein that show significant sequence similarity

```
NF1   841 TRATFMEVLTKILQQGTEFDTLAETVLADRFERLVELVTMMGDQGELPIA 890
Ira  1500 IRIAFLRVFIDIV...TNYPVNPEKHEMDKMLAIDDFLKYIIKNPILAFF 1546

      891 MALANVVPCSQWDELARVLVTLFDSRHLLYQLLWNMFSKEVELADSMQTL 940
     1547 GSLA..CSPADVDLYAGGFLNAFDTRNASHILVTELLKQEIKRAARSDDI 1594

      941 FRGNSLASKIMTFCFKVYGATYLQKLLDPLLRIVITSSDWQHVSFEVDPT 990
     1595 LRRNSCATRALSLYTRSRGNKYLIKTLRPVLQGIVDNKE....SFEID.. 1638

      991 RLEPSESLEENQRNLLQMTEKF....FHAIISSSSEFPPQLRSVCHCLYQ 1036
     1639 KMKPG...SENSEKMLDLFEKYMTRLIDAITSSIDDFPIELVDICKTIYN 1685

     1037 VVSQRFPQNSIGAVGSAMFLRFINPAIVSPYEAGILDKKPPPRIERGLKL 1086
     1686 AASVNFPEYAYIAVGSFVFLRFIGPALVSPDSENII.IVTHAHDRKPFIT 1734

     1087 MSKILQSIAN........HVLFTKEEHMRPFND....FVKSNFDAARRFF 1124
     1735 LAKVIQSLANGRENIFKKDILVSKEEFLKTCSDKIFNFLSELCKIPTNNF 1784

     1125 LDIASDCPTSDAVNHSL..............SFISDGNVLALHRLLWNN. 1159
     1785 TVNVREDPTPISFDYSFLHKFFYLNEFTIRKEIINESKLPGEFSFLKNTV 1834

     1160 ..QEKIGQYLSSNRDHKAVGRRPF.... DKMATLLAYLGPPEHKPVA 1200
     1835 MLNDKILGVLGQPSMEIKNEIPPFVVENREKYPSLYEFMSRYAFKKVD 1882
```