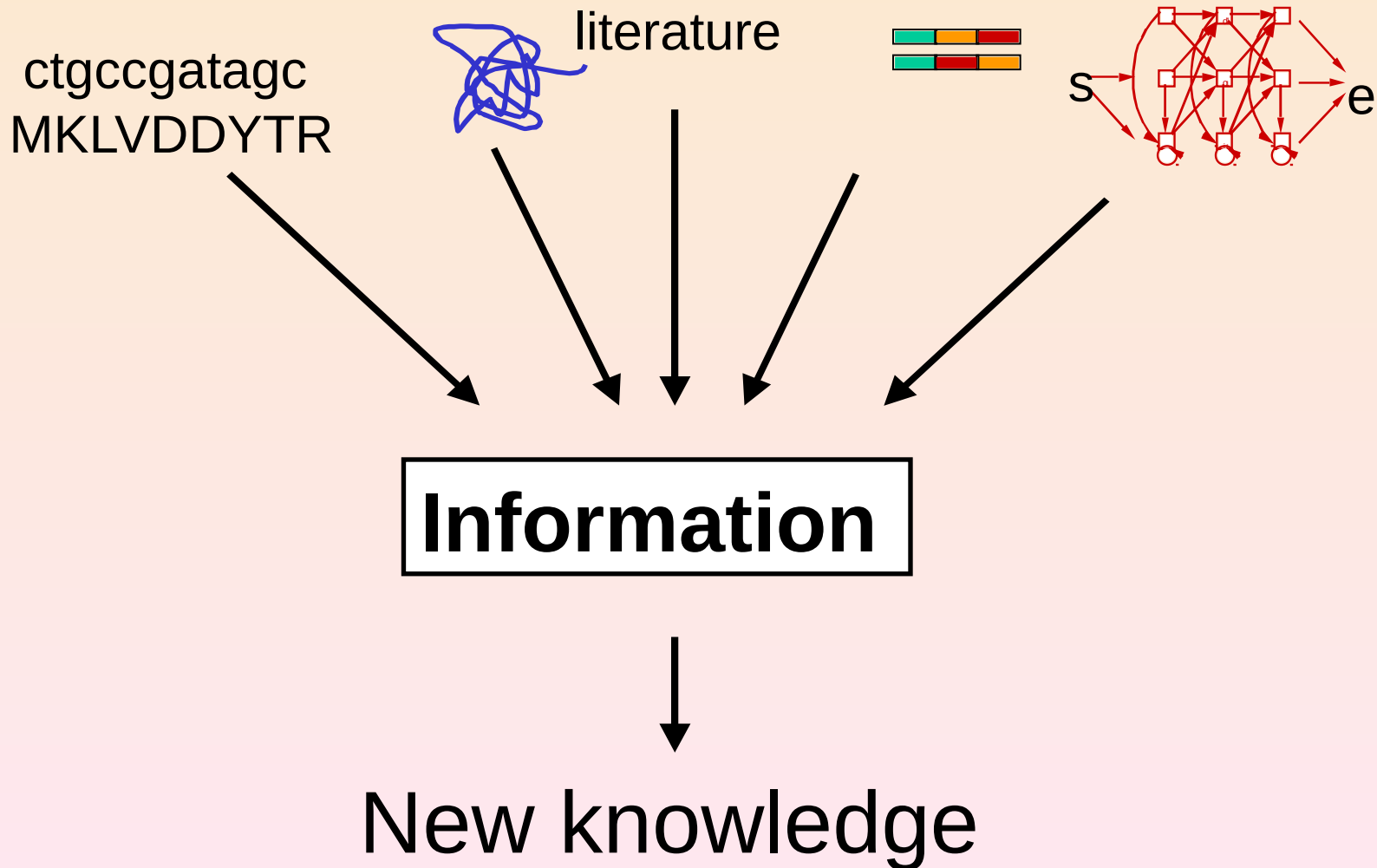


Biological Databases

G P S Raghava

Where do the data come from?

Example Databases



What is a Database/Resource?

NAR Database Issue (www.nar.oupjournals.org)

- **Collection of data in the related format**
 - structured
 - searchable (index) -> table of contents
 - updated periodically (release) -> new edition
 - cross-referenced ([hyperlinks](#)) -> links with other db
- Includes also associated tools (software) necessary for db access, db updating, db information insertion, db information deletion....
- **Type and Content of Data**
 - Sequence or Structure
 - Nucleic acid or protein
 - Important Biological information such as about enzyme and their metabolic pathways, mutations, diseases, drugs, images etc.
- **Based on source of data**
 - Primary database
 - Secondary database
 - Knowledge bases
 - Integrated Database

Database name	Direct sources of data			
	Other databases	Lit	LSSP	Sub
Primary				
GenBank (nuc.)	DDBJ, EMBL, GSDB	y	y	y
GenBank (prot.)	PIR, SWISS-PROT, PDB, PRF, Patents	y	n	y
EMBL	GenBank, DDBJ	y	y	y
DDBJ	GenBank, EMBL	y	y	y
GSDB	-	n	y	y
PIR	GenBank (translated), EMBL, DDBJ, MIPS (protein)	y	n	y
MIPS (prot.)	EMBL (translated)	y	y	n
MIPS (yeast)	-	n	y	y
SWISS-PROT	EMBL	y	n	y
PDB	-	y	n	y
YPD	MIPS, SGD	y	y	n
FlyBase	-	y	n	y
GDB	-	n	y	y
RHdb	-	n	y	y
MGD	Many, e.g. LocusBase and Matrix	y	n	n
RDP	GenBank, EMBL	y	n	y
RRNA SSU db	GenBank, EMBL	y	n	n
RRNA LSU db	GenBank, EMBL	y	n	n
p53 mutations	-	y	n	n
PROSITE	SWISS-PROT	y	n	n
Secondary				
TrEMBL	EMBL	n	n	n
ECD	GenBank, EMBL	y	n	n
NRSdb	GenBank, EMBL, DDBJ	n	n	n
SRPdb	GenBank, EMBL	y	n	y
PRINTS	OWL	n	n	n
BLOCK	PROSITE	n	n	n
HSSP	PDB, SWISS-PROT	n	n	n
FSSP	PDB	n	n	n
SBase	SWISS-PROT, PIR	y	n	n
TransTerm	GenBank	n	n	n
O-glycobase	SWISS-PROT	y	n	n
Knowledge bases				
SCOP	PDB	n	n	n
EMP collection	EMP (Enzymes and Metabolic Pathways database)	y	n	n
EcoCyc	EcoGene, SWISS-PROT	y	n	n
GIF db	-	y	n	n
CySPID	Many, e.g. GenBank, SWISS-PROT, FlyBase	y	n	y

Primary biological databases

- *Nucleic acid*

EMBL

GenBank

DDBJ (DNA Data Bank of
Japan)

- *Protein*

PIR

MIPS

SWISS-PROT

TrEMBL

NRL-3D

Nucleotide Databases

- **EMBL:** Nucleotide sequence database
- **Ensembl:** Automatics annotation of eukaryotic genomes
- **Genome Server:** Overview of completed genomes at EBI
- **Genome-MOT:** Genome monitoring table
- **EMBL-Align:** Multiple sequence alignment database
- **Parasites:** Parasite Genome databases
- **Mutations:** Sequence variation database project
- **IMGT:** Immunogenetics database, comprising-
IMGT/LIGM- database of immunoglobulins and
T-cell receptors, IMGT/HLA database of the human
MHC complex and IMGT/MHC covering MHC
complex of non-human species.

Reference site : www.ebi.ac.uk/Databases/nucleotide.html

EMBL/GenBank/DDJB

- These 3 db contain mainly the same information (few differences in the format and syntax)
- Serve as **archives** containing all sequences (single genes, ESTs, complete genomes, etc.) derived from:
 - Genome projects and sequencing centers
 - Individual scientists
 - Patent offices (i.e. USPTO, EPO)
- Non-confidential data are exchanged daily
- Currently: 2.5×10^7 sequences, over 3.2×10^{10} bp;
- Sequences from $> 50,000$ different species;

EMBL entry: example

ID HSERPG standard; DNA; HUM; 3398 BP.
XX
AC X02158;
XX
SV X02158.1
XX
DT 13-JUN-1985 (Rel. 06, Created)
DT 22-JUN-1993 (Rel. 36, Last updated, Version 2)
XX
DE Human gene for erythropoietin
XX
KW erythropoietin; glycoprotein hormone; hormone; signal peptide.
XX
OS Homo sapiens (human)
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN [1]
RP 1-3398
RX MEDLINE; 85137899.
RA Jacobs K., Shoemaker C., Rudersdorf R., Neill S.D., Kaufman R.J.,
RA Mufson A., Seehra J., Jones S.S., Hewick R., Fritsch E.F., Kawakita M.,
RA Shimizu T., Miyake T.;
RT Isolation and characterization of genomic and cDNA clones of human
RT erythropoietin;
RL Nature 313:806-810(1985).
XX
DR GDB; 119110; EPO.
DR GDB; 119615; TIMP1.
DR SWISS-PROT; P01588; EPO_HUMAN.
XX

keyword

taxonomy

references

Cross-references

...

EMBL entry (cont.)

```
CC Data kindly reviewed (24-FEB-1986) by K. Jacobs
FH Key Location/Qualifiers
FH
FT source 1..3398
FT /db_xref=taxon:9606
FT /organism=Homo sapiens
FT mRNA join(397..627,1194..1339,1596..1682,2294..2473,2608..3327)
FT CDS join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
FT /db_xref=SWISS-PROT:P01588
FT /product=erythropoietin
FT /protein_id=CAA26095.1
FT /translation=MGVHECPAWLWLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLLE
FT AKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAQVEVWQGLALLSEAVLRG
FT QALLVNSSQPWEPLQLHVDKAVSGLRSLTTLRALGAQKEAISPDAASAAPLRTITAD
FT TFRKLFrvysnflrgklklytgeacrtgdr
FT mat_peptide join(1262..1339,1596..1682,2294..2473,2608..2763)
FT /product=erythropoietin
FT sig_peptide join(615..627,1194..1261)
FT exon 397..627
FT /number=1
FT intron 628..1193
FT /number=1
FT exon 1194..1339
FT /number=2
FT intron 1340..1595
FT /number=2
FT exon 1596..1682
FT /number=3
FT intron 1683..2293
FT /number=3
FT exon 2294..2473
FT /number=4
FT intron 2474..2607
FT /number=4
FT exon 2608..3327
FT /note=3' untranslated region
FT /number=5
```

annotation

XX

sequence

```
SQ Sequence 3398 BP; 698 A; 1034 C; 991 G; 675 T; 0 other;
agcttctggg cttccagacc cagctacttt gcggaactca gcaaccagg catctctgag 60
tctccgccca agaccgggat gccccccagg aggtgtccgg gagcccagcc tttcccagat 120
```

GenBank file format

NCBI	Entrez	Nucleotide QUERY		BLAST	Entrez	?
------	--------	------------------	--	-------	--------	---

Other Formats:

Links:

LOCUS AB031330 1071 bp mRNA PRI 27-AUG-1999

DEFINITION Homo sapiens esp-1 mRNA for eosinophil serine protease, complete cds.

ACCESSION AB031330

NID g5777331

VERSION AB031330.1 GI:5777331

KEYWORDS eosinophil serine protease.

SOURCE Homo sapiens cell_line:HeLa S3 cDNA to mRNA, clone_lib:HeLa cDNA Lambda Triplex.

ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
Eutheria; Primates; Catarrhini; Hominidae; Homo.

GenBank file format

```
FEATURES                     Location/Qualifiers
    source                    1..1071
                              /organism="Homo sapiens"
                              /db_xref="taxon:9606"
                              /cell_line="HeLa S3"
                              /chromosome="16"
                              /clone_lib="HeLa cDNA Lambda TriplEx"
                              /map="16p13.3"
    gene                      4..948
    CDS                        4..948
                              /gene="esp-1"
                              /codon_start=1
                              /product="eosinophil serine protease"
                              /protein_id="BAA83521.1"
                              /db_xref="PID:d1047353"
                              /db_xref="PID:g5777332"
                              /db_xref="GI:5777332"
```

```
/translation="MGARGALLLALLARAGLRKPESQEAAPLSGPCGRRVITSRIVG
GEDAELGRWPWQGSRLRLWDSHVCVSLLSHRWALTAHCFETYSDLSDPSGWMVQFGQ
LTSMPFSFWSLQAYYTRYFVSNIIYLSPRYLGNSPYDIALVKLSAPVTYTKHIQPICLQA
STFEFENRTDCWVTGWGYIKEDALPSPHTLQEVQVAIINNSMCNHLFLKYSFRKIDF
GDMVCAGNAQGGKDACFGDSGGPLACNKNGLWYQIGVVSWGVGCGRPNRPGVYTNISH
HFEWIKQLMAQSGMSQPDPSWPLLFFPLLWALPLLGPV"
```

```
1058
/note="13 a nucleotides"
```

```
BASE COUNT      208 a    317 c    305 g    241 t
ORIGIN
```

```

1 gccatggggcg cgcgcggggc gctgctgctg gcgctgctgc tggctcgggc tggactcagg
61 aagccggagtg cgcaggaggc ggcgcgcgta tcaggaccat ggcggccgacg ggtcatcacg
121 tcgcgcatcg tgggtggaga ggacgccgaa ctccggcggt ggccgtggca ggggagcctg
181 cgccctgtggg attccacgt atgcggagtg agcctgctca gccaccgctg ggcactcacg
241 gcggcgcaact gctttgaaac ctatagtgac cttagtgatc cctccgggtg gatggtccag
301 tttggccagc tgacttccat gccatccttc tggagcctgc aggcctacta caccgcgttac
361 ttcgtatcga atatctatct gagccctcgc tacctgggga attcacccta tgacattgcc
421 ttggtgaagc tgtctgcacc tgtcacctac actaaacaca tccagcccat ctgtctccag
```

Databases related to Genomics

- Contain information on genes, gene location (mapping), gene nomenclature and links to sequence databases;
- Exist for most organisms important for life science research;
- Examples: MIM, GDB (human), MGD (mouse), FlyBase (Drosophila), SGD (yeast), MaizeDB (maize), SubtiList (B.subtilis), etc.
- Format: generally relational (Oracle, SyBase or AceDb).



PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Help

Search for on chromosome(s) Find

☐ Show linked entries [Help](#) [FTP](#) ☐ Advanced search

[Entrez Genomes](#)

[Prominent organisms](#)

[Maps](#)

Map Viewer Help

Human Maps Help

Mouse Maps Help

Human/Mouse Homology Map

[Related Resources](#)

Human Genome Guide

Mouse Genome Guide

LocusLink

OMIM

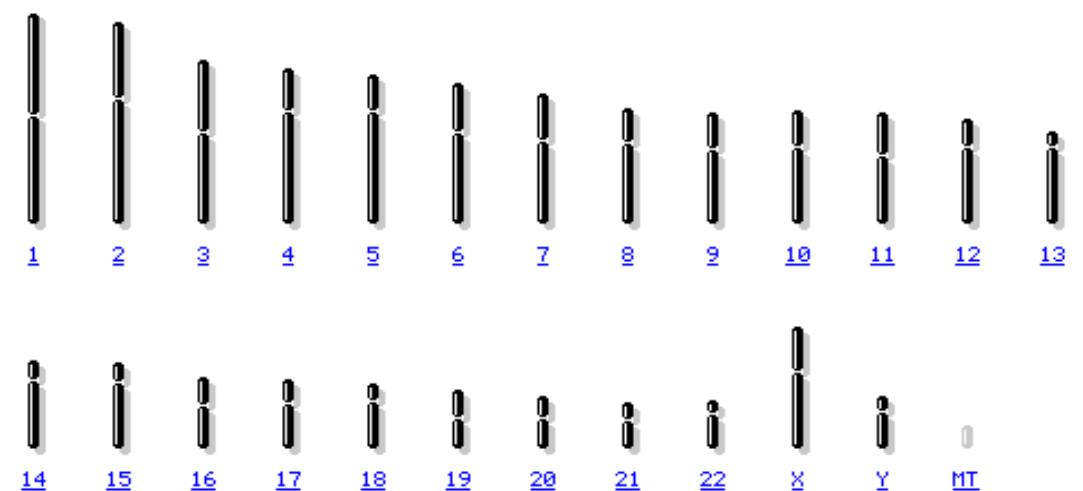
UniGene

[Sequence Data](#)

Human Genome

[Homo sapiens genome view](#) [build 28](#)

[BLAST search the human genome](#)



1 2 3 4 5 6 7 8 9 10 11 12 13

14 15 16 17 18 19 20 21 22 X Y MT

The NCBI Map Viewer provides graphical displays of features on NCBI's assembly of human genomic sequence data as well as cytogenetic, genetic, physical, and radiation hybrid maps. [Release notes](#) report changes in MapView displays or

[About Entrez](#)

[Entrez Genomes](#)
Help

[Submitting genome sequences](#)

[All Organisms](#)

[Prominent Organisms](#)

[Microbial genomes](#)
[Taxonomy Tree](#)
[BLAST](#)
[List of projects](#)
[PDB neighbors](#)

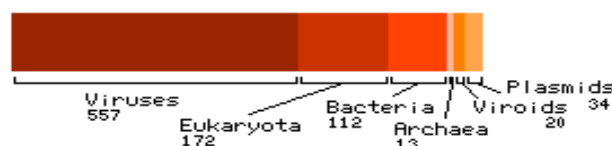
[Archaea](#)
Genome
Plasmids

[Bacteria](#)
Genome
Plasmids

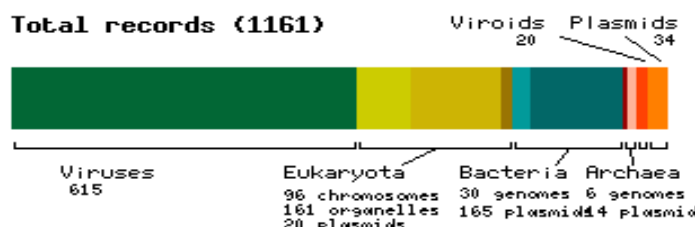
The whole genomes of over 800 organisms can be found in Entrez Genomes. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life - [bacteria](#), [archaea](#), and [eukaryota](#) - are represented, as well as many [viruses](#) and [organelles](#).

Entrez Genomes statistics

Total species (908)



Total records (1161)



Latest complete genome: *Halobacterium sp.* NRC-1

Publication date: October 3 2000
Size: 2,014,239 bp

Reference: Na, W.Y., et al. (2000) Genome

Related Resources

[Homo sapiens](#)
Map Viewer
[Drosophila melanogaster](#)
sequence and resources
[Microbial](#)
sequencing projects list
[Organelle](#)
reference sequences and tools

[Malaria](#)
genetics and genomics
[Retrovirus](#)
tools and resources

[BLAST](#)
unfinished microbial genomes
[COGs](#)
clusters of orthologous groups

Major Sequencing Centers

Ensembl

- Contains all the human genome DNA sequences currently available in the public domain.
- Automated annotation: by using different software tools, features are identified in the DNA sequences:
 - Genes (known or predicted)
 - Single nucleotide polymorphisms (SNPs)
 - Repeats
 - Homologies
- Created and maintained by the EBI and the Sanger Center (UK)
- www.ensembl.org



Zebrafish

Flat File FASTA **Gene List** Feature List SNP List Image

Select genes to export

Region

Chromosome:

12

☐ Bands from:

to ---

☐ Contigs from:

to

☐ Markers from:

to

☒ Bases from:

72000000

to

85000000

☐ Entire Chromosome

☐ Entire Genome

Restrict selection

Include:

☐ Known genes only

☐ Exclude Known genes

☒ Both

Include:

☐ Disease genes only

☐ Exclude Disease genes

☒ Both

Include:

☐ Transmembrane domains only

☐ Exclude Transmembrane domains

☒ Both

Include:

☐ Low-complexity domains only

☐ Exclude Low-complexity domains

☒ Both

Include:

☐ Signal domains only

☐ Exclude Signal domains

☒ Both

Include only those genes with

Family

ID:

Protein Databases

- **SWISS-PROT:** Annotated Sequence Database
- **TrEMBL:** Database of EMBL nucleotide translated sequences
- **InterPro:** Integrated resource for protein families, domains and functional sites.
- **CluSTr:** Offers an automatic classification of SWISS-PROT and TrEMBL.
- **IPI:** A non-redundant human proteome set constructed from SWISS-PROT, TrEMBL, Ensembl and RefSeq.
- **GOA:** Provides assignments of gene products to the Gene Ontology (GO) resource.
- **Proteome Analysis:** Statistical and comparative analysis of the predicted proteomes of fully sequenced organisms
- **Protein Profiles:** Tables of SWISS-PROT and TrEMBL entries and alignments for the protein families of the Protein Profile.
- **IntEnz:** The Integrated relational Enzyme database (IntEnz) will contain enzyme data approved by the Nomenclature Committee.

Reference site : www.ebi.ac.uk/Databases/protein.html

Swiss-Prot

- Annotated protein sequence database established in 1986 and maintained collaboratively since 1987, by the Department of Medical Biochemistry of the University of Geneva and EBI
- Complete, Curated, Non-redundant and cross-referenced with 34 other databases
- Highly cross-referenced
- Available from a variety of servers and through sequence analysis software tools
- More than 8,000 different species
- First 20 species represent about 42% of all sequences in the database
- More than 1,29,000 entries with 4.7×10^{10} amino acids
- More than 6,22,000 entries in TrEMBL

TrEMBL (Translation of EMBL)

- Computer-annotated supplement to SWISS-PROT, as it is impossible to cope with the flow of data...
- Well-structure SWISS-PROT-like resource
- Derived from automated EMBL CDS translation maintained at the EBI, UK.
- TrEMBL is automatically generated and annotated using software tools (incompatible with the SWISS-PROT in terms of quality)
- TrEMBL contains all what is **not yet** in SWISS-PROT

SWISS-PROT file format

General information about the entry	
Entry name	FA12_HUMAN
Primary accession number	P00748
Secondary accession number(s)	None
Entered in SWISS-PROT in	Release 01, July 1986
Sequence was last modified in	Release 12, October 1989
Annotations were last modified in	Release 35, November 1997
Name and origin of the protein	
Protein name	COAGULATION FACTOR XII [Precursor]
Synonym(s)	EC 3.4.21.38 HAGEMAN FACTOR HAF
Gene name(s)	F12
From	Homo sapiens (Human)
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

SWISS-PROT file format

Comments

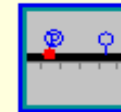
- **FUNCTION:** FACTOR XII IS A SERUM GLYCOPROTEIN THAT PARTICIPATES IN THE INITIATION OF BLOOD COAGULATION, FIBRINOLYSIS, AND THE GENERATION OF BRADYKININ AND ANGIOTENSIN.
- **CATALYTIC ACTIVITY:** CLEAVES SELECTIVELY ARG-|-ILE BONDS AND ACTIVATES COAGULATION FACTORS VII AND XI.
- **PTM:** O- AND N-GLYCOSYLATED.
- **DISEASE:** DEFECTS IN F12 DO NOT CAUSE ANY CLINICAL SYMPTOMS. THE SOLE EFFECT IS THAT WHOLE-BLOOD CLOTTING TIME IS PROLONGED.
- **MISCELLANEOUS:** FACTOR XII, PREKALLIKREIN, AND HMW KININOGEN FORM A COMPLEX BOUND TO AN ANIONIC SURFACE. PREKALLIKREIN IS CLEAVED BY FACTOR XII TO FORM KALLIKREIN, WHICH THEN CLEAVES FACTOR XII FIRST TO ALPHA-FACTOR XIIA AND THEN TO BETA-FACTOR XIIA. ALPHA-FACTOR XIIA ACTIVATES FACTOR XI TO FACTOR XIA.
- **SIMILARITY:** CONTAINS 2 EGF-LIKE DOMAINS.
- **SIMILARITY:** CONTAINS 1 FIBRONECTIN TYPE-I DOMAIN.
- **SIMILARITY:** CONTAINS 1 FIBRONECTIN TYPE-II DOMAIN.
- **SIMILARITY:** CONTAINS 1 KRINGLE REGION.
- **SIMILARITY:** BELONGS TO [PEPTIDASE](#) FAMILY S1; ALSO KNOWN AS THE TRYPSIN FAMILY.

SWISS-PROT file format

Cross-references	
EMBL	M31315; AAA70225.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M11723; AAA51986.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M17466; AAB59490.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M17464; AAB59490.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] M17465; AAB59490.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] M13147; AAA70224.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A29411; KFHU12.
HSSP	P00763 ; 1DPO. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
MIM	234000 ; -.
GeneCards	GeneCards ; F12.
PFAM	PF00008 ; EGF; 2. PF00039 ; fn1; 1. PF00040 ; fn2; 1. PF00051 ; kringle; 1. PF00089 ; trypsin; 1.
	PS00021 ; KRINGLE 1; 1. PS00022 ; EGF 1; 2.

SWISS-PROT file format

DOMAIN	217	295	KRINGLE.
DOMAIN	296	349	PRO-RICH.
DOMAIN	373	615	CATALYTIC.
CARBOHYD	109	109	FUCOSE.
CARBOHYD	249	249	
CARBOHYD	299	299	POTENTIAL.
CARBOHYD	305	305	POTENTIAL.
CARBOHYD	308	308	POTENTIAL.
CARBOHYD	328	328	POTENTIAL.
CARBOHYD	329	329	POTENTIAL.
CARBOHYD	337	337	POTENTIAL.
ACT_SITE	412	412	CHARGE RELAY SYSTEM (BY SIMILARITY).
ACT_SITE	461	461	CHARGE RELAY SYSTEM (BY SIMILARITY).
ACT_SITE	563	563	CHARGE RELAY SYSTEM (BY SIMILARITY).
DISULFID	98	110	BY SIMILARITY.
DISULFID	104	119	BY SIMILARITY.
DISULFID	121	130	BY SIMILARITY.



[FT table viewer](#)

Sequence information

Length: **615 AA** [This is the length of the unprocessed precursor]

Molecular weight: **67818 Da** [This is the Mw of the unprocessed precursor]

CRC32: **282B2A6B** [This is a checksum on the sequence]

10	20	30	40	50	60
MRALLLLGFL	LVSLESTLSI	PPWEAPKEHK	YKAEHTVVL	TVTGEPCHP	FQYHRQLYHK
70	80	90	100	110	120
CTHKGRPGPQ	PWCATTPNFD	QDQRWGYCLE	PKKVVDHCSK	HSPCQKGGTC	VNMPSGPHCL
130	140	150	160	170	180
CPQHLTGNHC	QKEKCFEPQL	LRFFHKNEIW	YRTEQAAVAR	CQCKGPDHC	QRLASQACRT

Structure Databases

- **MSD:** The Macromolecular Structure Database –
A relational database representation of clean Protein Data Bank (PDB)
- **3DSeq:** 3D sequence alignment server- Annotation of the alignments between sequence database and the PDB
- **FSSP:** Based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB)
- **DALI:** Fold Classification based on Structure-Structure Assignments
- **3Dee:** Database of protein domain definitions wherein the domains have been clustered on sequence and structural similarity
- **NDB:** Nucleic Acid Structure Database

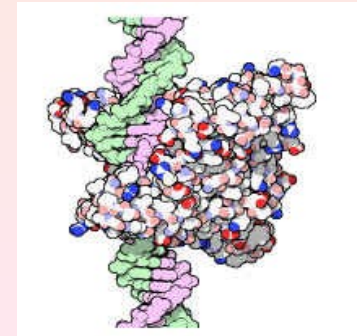
Selected WWW database resources for macromolecular structures.

Databases	URL
Structure and sequence/structure databases	
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/
FSSP	http://www2.ebi.ac.uk/dali/fssp/
Molecular Modeling Database	http://www.ncbi.nlm.nih.gov/Structure/
CAMPASS	http://www-cryst.bioc.cam.ac.uk/~campass/
ISSD	http://www.protein.bio.msu.su/issd/
Library of Protein Family Cores (LPFC)	http://WWW-SMI.Stanford.EDU/projects/helix/LPFC/
3D_ALI (a database of aligned protein structures and related sequences)	http://www.embl-heidelberg.de/argos/ali/ali_info.html
IDITIS (relational database and query tool for proteins)	http://www.oxmol.co.uk/prods/identis/
HSSP	http://www.sander.embl-heidelberg.de/hssp/
Speciality databases	
HIV Protease Database	http://www-lbse.ncifcrf.gov/HIVdb/
Nucleic Acid Database	http://ndbserver.rutgers.edu/
Prolysis (protease and protease inhibitor Web server)	http://delphi.phys.univ-tours.fr/Prolysis/
International Immunogenetics Database (IMGT)	http://imgt.cnusc.fr:8104/
Enzyme Structures Database	http://www.biochem.ucl.ac.uk/bsm/enzymes/
Features databases	
Molecular Movements Database	http://bioinfo.mbb.yale.edu/MolMovDB/
OLDERADO	http://neon.chem.le.ac.uk/olderado/
PROCAT	http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html
Protein Quaternary Structures (PQS)	http://pqs.ebi.ac.uk/
ReLiBase (receptor-ligand complexes database)	http://www2.ebi.ac.uk:8081/home.html
PROMISE	http://bioinf.leeds.ac.uk/promise/
PDBSum	http://www.biochem.ucl.ac.uk/bsm/pdbsum/
Biological Macromolecule Crystallization Database (BMCD)	http://h178133.nist.gov:4400/bmcd/bmcd.html
Resources	
Protein Data Bank	http://www.rcsb.org/pdb/

Protein DataBank (PDB)

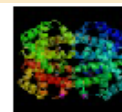
- Important in solving real problems in molecular biology
- Protein Databank
 - PDB Established in 1972 at Brookhaven National Laboratory (BNL)
 - Sole international repository of macromolecular structure data
 - Moved to Research Collaboratory for Structural Bioinformatics

<http://www.rcsb.org/>



Effective use of PDB

- Queries are of three types
 - PDBid - As quoted in paper
 - Search Lite - one or more keywords
 - Search Fields - A detailed query form
- Query results
 - Structure Explorer - details of the structure
 - Query Result Browser - for multiple structures
- PDB Viewer



Summary Information



Summary Information

[View Structure](#)

[Download/Display File](#)

[Structural Neighbors](#)

[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

[Crystallization Info](#)

[Previous version\(s\):
1HHB](#)

Explore

[SearchLite](#) [SearchFields](#)

Compound: **Hemoglobin (Deoxy)**

Authors: **G. Fermi, M. F. Perutz**

Exp. Method: **X-ray Diffraction**

Classification: **Oxygen Transport**

Source: **Homo Sapiens**

Primary Citation: **Fermi, G., Perutz, M. F., Shaanan, B., Fourme, R.: The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol* 175 pp. 159 (1984)**
[[Medline](#)]

Deposition Date: **07-Mar-1984**

Release Date: **17-Jul-1984**

Resolution [Å]: **1.74**

R-Value: **0.135**

Space Group: **P 21**

Unit Cell: *dim [Å]:* *a* **63.15** *b* **83.59** *c* **53.80**

angles [°]: *alpha* **90.00** *beta* **99.34** *gamma* **90.00**

Polymer Chains: **A, B, C, D**

Residues: **574**

Atoms: **4779**

HET groups:

ID	Name	Formula
HEM	PROTOPORPHYRIN IX CONTAINING FE	C ₃₄ H ₃₂ N ₄ O ₄ FE ₁
PO4	PHOSPHATE ION	O ₄ P ₁

Other Versions: [2HHB](#), [3HHB](#)

PDB: example

```
HEADER  LYASE(OXO-ACID)                01-OCT-91  12CA  12CA  2
COMPND  CARBONIC ANHYDRASE /II (CARBONATE DEHYDRATASE) (/HCA II)  12CA  3
COMPND  2 (E.C.4.2.1.1) MUTANT WITH VAL 121 REPLACED BY ALA (/V121A) 12CA  4
SOURCE  HUMAN (HOMO SAPIENS) RECOMBINANT PROTEIN                12CA  5
AUTHOR  S.K.NAIR,D.W.CHRISTIANSON                      12CA  6
REVDAT  1 15-OCT-92 12CA  0                      12CA  7
JRNL    AUTH  S.K.NAIR,T.L.CALDERONE,D.W.CHRISTIANSON,C.A.FIERKE 12CA  8
JRNL    TITL  ALTERING THE MOUTH OF A HYDROPHOBIC POCKET.        12CA  9
JRNL    TITL  2 STRUCTURE AND KINETICS OF HUMAN CARBONIC ANHYDRASE 12CA 10
JRNL    TITL  3 /II$ MUTANTS AT RESIDUE VAL-121                12CA 11
JRNL    REF   J.BIOL.CHEM.          V. 266 17320 1991    12CA 12
JRNL    REFN  ASTM JBCHA3  US ISSN 0021-9258            071 12CA 13
REMARK  1                                12CA 14
REMARK  2                                12CA 15
REMARK  2 RESOLUTION. 2.4  ANGSTROMS.                12CA 16
REMARK  3                                12CA 17
REMARK  3 REFINEMENT.                                12CA 18
REMARK  3 PROGRAM          PROLSQ                    12CA 19
REMARK  3 AUTHORS          HENDRICKSON,KONNERT        12CA 20
REMARK  3 R VALUE          0.170                    12CA 21
REMARK  3 RMSD BOND DISTANCES  0.011  ANGSTROMS        12CA 22
REMARK  3 RMSD BOND ANGLES    1.3  DEGREES            12CA 23
REMARK  4                                12CA 24
REMARK  4 N-TERMINAL RESIDUES SER 2, HIS 3, HIS 4 AND C-TERMINAL 12CA 25
REMARK  4 RESIDUE LYS 260 WERE NOT LOCATED IN THE DENSITY MAPS AND, 12CA 26
REMARK  4 THEREFORE, NO COORDINATES ARE INCLUDED FOR THESE RESIDUES. 12CA 27
.....
```

PDB (cont.)

```
SHEET 3 S10 PHE 66 PHE 70 -1 O ASN 67 N LEU 60 12CA 68
SHEET 4 S10 TYR 88 TRP 97 -1 O PHE 93 N VAL 68 12CA 69
SHEET 5 S10 ALA 116 ASN 124 -1 O HIS 119 N HIS 94 12CA 70
SHEET 6 S10 LEU 141 VAL 150 -1 O LEU 144 N LEU 120 12CA 71
SHEET 7 S10 VAL 207 LEU 212 1 O ILE 210 N GLY 145 12CA 72
SHEET 8 S10 TYR 191 GLY 196 -1 O TRP 192 N VAL 211 12CA 73
SHEET 9 S10 LYS 257 ALA 258 -1 O LYS 257 N THR 193 12CA 74
SHEET 10 S10 LYS 39 TYR 40 1 O LYS 39 N ALA 258 12CA 75
TURN 1 T1 GLN 28 VAL 31 TYPE VIB (CIS-PRO 30) 12CA 76
TURN 2 T2 GLY 81 LEU 84 TYPE II(PRIME) (GLY 82) 12CA 77
TURN 3 T3 ALA 134 GLN 137 TYPE I (GLN 136) 12CA 78
TURN 4 T4 GLN 137 GLY 140 TYPE I (ASP 139) 12CA 79
TURN 5 T5 THR 200 LEU 203 TYPE VIA (CIS-PRO 202) 12CA 80
TURN 6 T6 GLY 233 GLU 236 TYPE II (GLY 235) 12CA 81
CRYST1 42.700 41.700 73.000 90.00 104.60 90.00 P 21 2 12CA 82
ORIGX1 1.000000 0.000000 0.000000 0.00000 12CA 83
ORIGX2 0.000000 1.000000 0.000000 0.00000 12CA 84
ORIGX3 0.000000 0.000000 1.000000 0.00000 12CA 85
SCALE1 0.023419 0.000000 0.006100 0.00000 12CA 86
SCALE2 0.000000 0.023981 0.000000 0.00000 12CA 87
SCALE3 0.000000 0.000000 0.014156 0.00000 12CA 88
ATOM 1 N TRP 5 8.519 -0.751 10.738 1.00 13.37 12CA 89
ATOM 2 CA TRP 5 7.743 -1.668 11.585 1.00 13.42 12CA 90
ATOM 3 C TRP 5 6.786 -2.502 10.667 1.00 13.47 12CA 91
ATOM 4 O TRP 5 6.422 -2.085 9.607 1.00 13.57 12CA 92
ATOM 5 CB TRP 5 6.997 -0.917 12.645 1.00 13.34 12CA 93
ATOM 6 CG TRP 5 5.784 -0.209 12.221 1.00 13.40 12CA 94
ATOM 7 CD1 TRP 5 5.681 1.084 11.797 1.00 13.29 12CA 95
ATOM 8 CD2 TRP 5 4.417 -0.667 12.221 1.00 13.34 12CA 96
ATOM 9 NE1 TRP 5 4.388 1.418 11.515 1.00 13.30 12CA 97
ATOM 10 CE2 TRP 5 3.588 0.375 11.797 1.00 13.35 12CA 98
ATOM 11 CE3 TRP 5 3.837 -1.877 12.645 1.00 13.39 12CA 99
ATOM 12 CZ2 TRP 5 2.216 0.208 11.656 1.00 13.39 12CA 100
ATOM 13 CZ3 TRP 5 2.465 -2.043 12.504 1.00 13.33 12CA 101
ATOM 14 CH2 TRP 5 1.654 -1.001 12.009 1.00 13.34 12CA 102
```

.....

A comprehensive
help document

Cn3D FAQ

Frequently Asked
Questions

Cn3D Install

Installation and
Configuration

MMDB

NCBI's structure
database

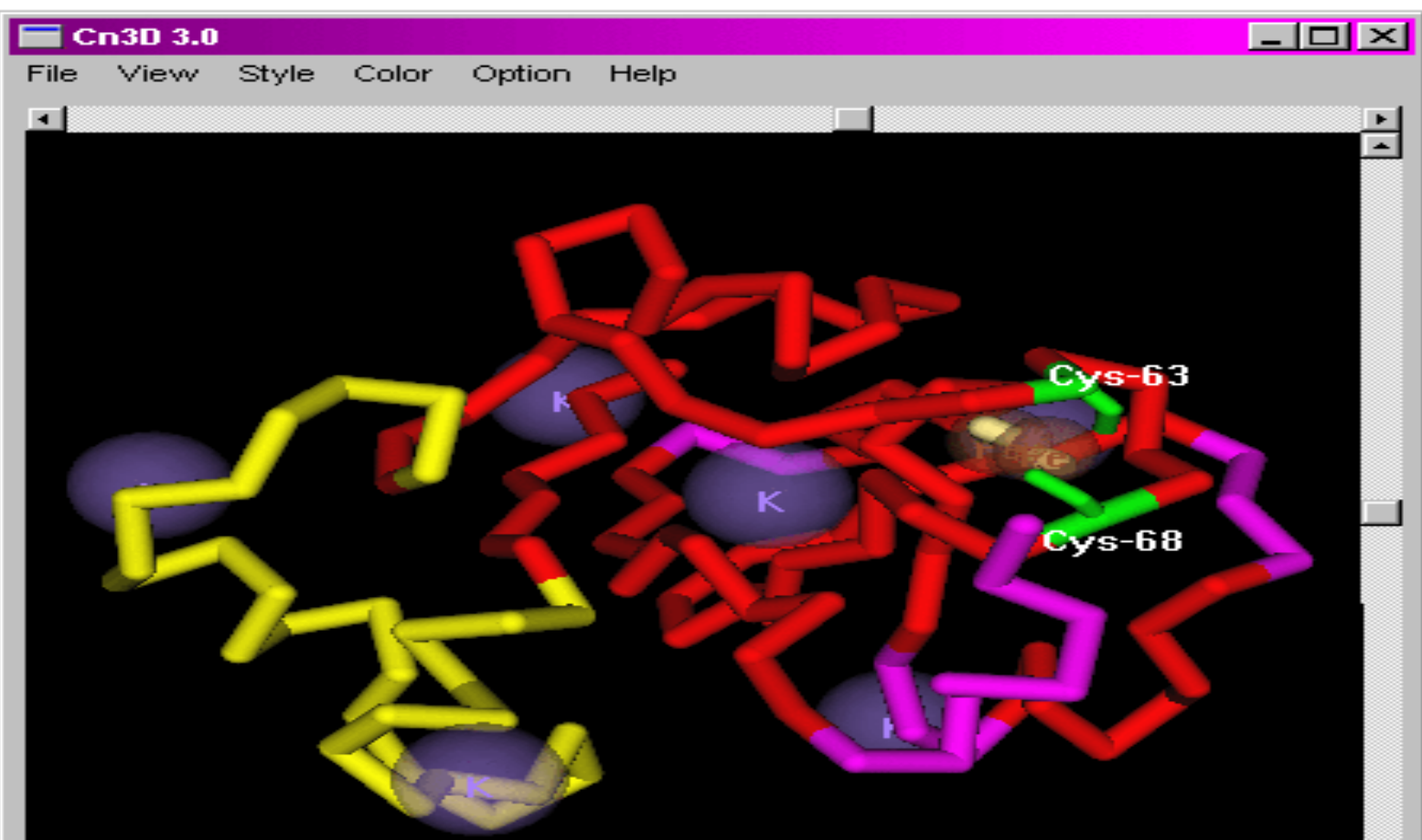
PDBeast

Taxonomy in
MMDB

VAST

Structure
comparisons

VAST Search

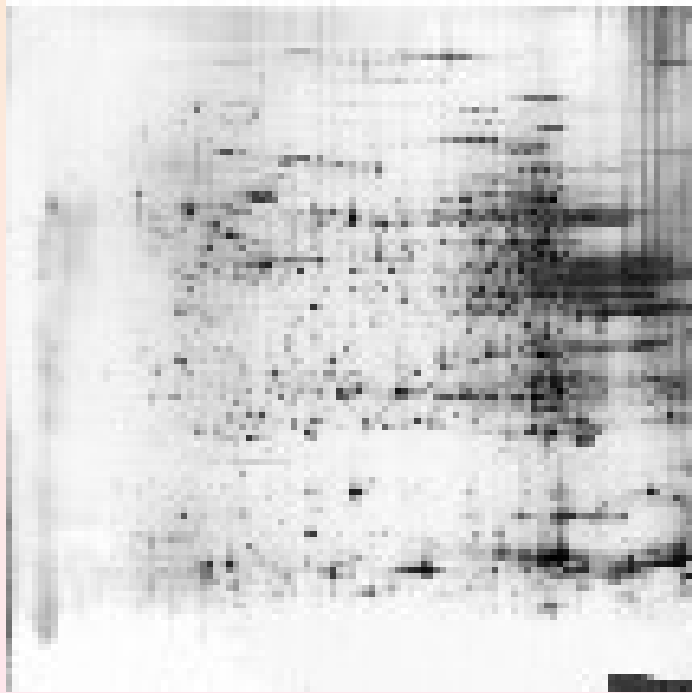


Databases related to Proteomics

- Contain information obtained by 2D-PAGE: master images of the gels and description of identified proteins
- Examples: SWISS-2DPAGE, ECO2DBASE, Maize-2DPAGE, Sub2D, Cyano2DBase, etc.
- Format: composed of image and text files
- Most 2D-PAGE databases are “federated” and use SWISS-PROT as a master index
- Mass Spectrometry (MS) database

Proteomics

1978-1998



MALDI-TOF?
ESI-MS?

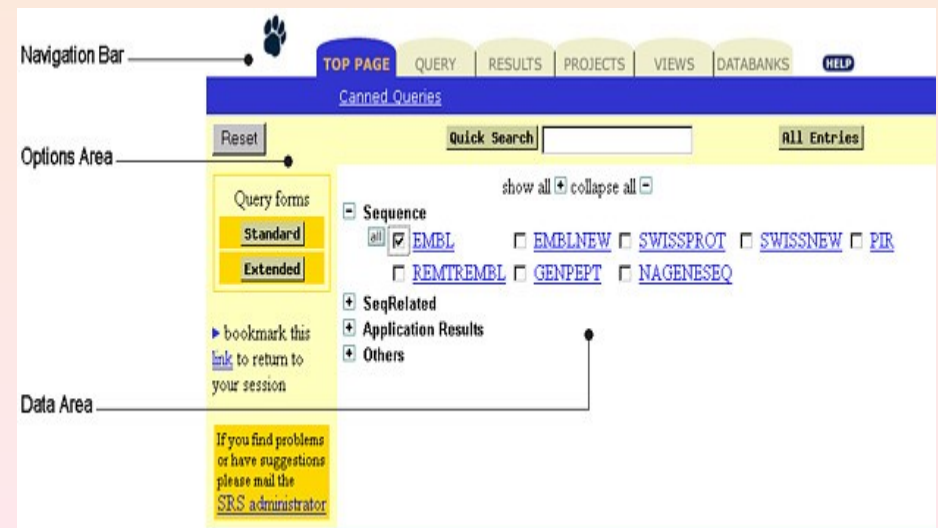
Database Mining Tools

- **SRS**: Sequence Retrieval System
- **Entrez**: Search Engine at NCBI, US
- **Bankit**: World Wide Web sequence submission server
- Sequence Similarity Search Tools-**BLAST & FASTA**
 - Finding sequence homologs to deduce the identity of query sequence
 - Identify potential sequence homologs with known three dimensional structure

Sequence Retrieval System

SRS is a powerful data integration platform

- Provides rapid, easy and user friendly access
- Large volumes of heterogeneous Life Science data
- Stored in more than 400 internal and public domain databases
- Available at <http://srs.ebi.ac.uk/>



....SRS

SRS performs searches on the following categories:

References	Sequence libraries - complete
Sequence libraries - subsections	InterPro&Related
SeqRelated	TransFac
User Owned Databanks	Application Results
Protein3DStruct	Genome
Mapping	Mutations
Locus Specific Mutations	Metabolic Pathways
Others	SNP
EMBOSS DOCS	System

Searches can be carried out using

- Quick search on all entries
- Standard form with Boolean operators
- Extended form with field names

Entrez at NCBI

It is a retrieval system for searching several linked databases such as

- PubMed: The biomedical literature (PubMed)
- Nucleotide sequence database (Genbank)
- Protein sequence database
- Structure: Three-dimensional macromolecular structures
- Genome: Complete genome assemblies
- PopSet: Population study data sets
- OMIM: Online Mendelian Inheritance in Man
- Taxonomy: Organisms in GenBank
- Books: Online books
- ProbeSet: Gene expression and microarray datasets
- 3D Domains: Domains from Entrez Structure
- UniSTS: Markers and mapping data
- SNP: Single nucleotide polymorphisms
- CDD: Conserved domains

Entrez: Search fields

- Keyword allows to search a set of indexed terms
- Accession allows to search accession numbers
- Author Name
- Affiliations of authors
- Journal Title
- E.C. Numbers
- Feature Key searches for particular DNA feature
- SeqId is string identifier
- Title Words
- Text Words
- Organism
- Pubmed ID
- Publication and modification date
- Protein Name

Search PubMed for HMM Go Clear

Limits Preview/Index History Clipboard

Display Summary Save Text Order Details Add to Clipboard

Show: 20 Items 1-20 of 991 Page 1 of 50 Select page: 1 2 3 4 5 6 7 8 9 10 >>

1: [Joel PB, Trybus KM, Sweeney HL.](#)
[Related Articles](#)

Two conserved lysines at the 50/20 kDa junction of myosin are necessary for triggering actin-activation.

J Biol Chem. 2000 Oct 20 [epub ahead of print]

[Record as supplied by publisher]

PMID: 11042210

2: [Qin F, Auerbach A, Sachs F.](#)
[Related Articles](#)

Hidden markov modeling for single channel kinetics with filtering and correlated noise.

Biophys J. 2000 Oct;79(4):1928-44.

[MEDLINE record in process]

PMID: 11023898; UI: 20480155

3: [Qin F, Auerbach A, Sachs F.](#)
[Related Articles](#)

A direct optimization approach to hidden markov modeling for single channel kinetics.

Biophys J. 2000 Oct;79(4):1915-27.

[MEDLINE record in process]

PMID: 11023897; UI: 20480154

4: [Sweeney HL, Chen LQ, Trybus KM.](#)
[Related Articles](#)

Regulation of asymmetric smooth muscle myosin II molecules.

J Biol Chem. 2000 Oct 3 [epub ahead of print]

[Record as supplied by publisher]

PMID: 11018047

[About Entrez](#)

Entrez PubMed

[Overview](#)
[Help | FAQ](#)
[New/Noteworthy](#)

PubMed Services

[Journal Browser](#)
[MeSH Browser](#)
[Single Citation Matcher](#)
[Batch Citation Matcher](#)
[Clinical Queries](#)
[Cubby !\[\]\(44520364e8864bead6d6828e9db2aebc_img.jpg\)](#)

Related Resources

[Order Documents](#)
[Grateful Med](#)
[Consumer Health](#)
[Clinical Alerts](#)
[ClinicalTrials.gov](#)

[Privacy Policy](#)

- 1: [AAG25673](#)

alpha-D chain hemoglobin [*Geochelone carbonaria*]

gi|10953950|gb|AAG25673.1|AF304335_1[10953950]

[Nucleotide](#), [Taxonomy](#)
- 2: [AAA97887](#)

nonsymbiotic hemoglobin [*Glycine max*]

gi|1276977|gb|AAA97887.1|[1276977]

[PubMed](#), [Related Sequences](#), [Nucleotide](#), [Taxonomy](#)
- 3: [1DLYA](#)

Chain A, X-Ray Crystal Structure Of Hemoglobin From The Green Unicellular Alga *Chlamydomonas Eugametos*

gi|10835657|pdb|1DLY|A[10835657]

[PubMed](#), [Related Sequences](#), [Structure](#), [Taxonomy](#)
- 4: [1DLWA](#)

Chain A, X-Ray Crystal Structure Of Truncated Hemoglobin From *P. Caudatum*.

gi|10835656|pdb|1DLW|A[10835656]

[PubMed](#), [Related Sequences](#), [Structure](#), [Taxonomy](#)
- 5: [AAF04173](#)

class 2 non-symbiotic hemoglobin; 69592-70841 [*Arabidopsis thaliana*]

gi|6119529|gb|AAF04173.1|AC011560_14[6119529]

[Related Sequences](#), [Nucleotide](#), [Taxonomy](#)
- 6: [AAG22831](#)

hemoglobin [*Ceratodon purpureus*]

gi|10764841|gb|AAG22831.1|AF309562_1[10764841]

[Related Sequences](#), [Nucleotide](#), [Taxonomy](#)

File Formats of the sequences

Readseq (<http://bimas.dcrct.nih.gov/molbio/readseq/>)

- | | |
|---------------------------|------------------------------|
| 1. IG/Stanford | 10. Olsen (in-only) |
| 2. GenBank/GB | 11. Phylip3.2 |
| 3. NBRF | 12. Phylip |
| 4. EMBL | 13. Plain/Raw |
| 5. GCG | 14. PIR/CODATA |
| 6. DNASTrider | 15. MSF |
| 7. Fitch | 16. ASN.1 |
| 8. Pearson/Fasta | 17. PAUP |
| 9. Zuker (in-only) | 18. Pretty (out-only) |

FAST Format

- Popular Format and commonly used

> Seq1

ALVLRARLATGPATGCTRTARARLATGALVLRARLATGPARARLATGPATGCTRTARA
RLATGALVLRARRLATGPATGCTRRLATGPATGCTRRARLATGPATGCTRTARARLAT
GALVLRAR

>Seq2

TGCTRTARARLATGALVLRARLATGPARARALVLRARLATGPATGCTRTARATGALVL
RARLATGPARARALVLRARLATG

>Seq 3

.....

Intelligenetics format

```
;seq1, 16 bases, 2688 checksum.  
seq1  
agctagctagctagct1  
;seq2, 16 bases, 25C8 checksum.  
seq2  
aactaactaactaact1
```

NBRF format

```
>DL;seq1  
seq1, 16 bases, 2688 checksum.  
agctagctag ctagct*
```

```
>DL;seq2  
seq2, 16 bases, 25C8 checksum.  
aactaactaa ctaact*
```

GCG format

```
seq1
    seq1  Length: 16  Check: 9864  ..
    1  agctagctag ctagct

seq2
    seq2  Length: 16  Check: 9672  ..
    1  aactaactaa ctaact
```

GCG multiple sequence format (MSF)

```
/tmp/readseq.in.2449  MSF: 16 Type: N January 01,  
1776 12:00 Check: 9536 ..
```

```
  Name: seq1                      Len:      16 Check:   9864  
Weight: 1.00  
  Name: seq2                      Len:      16 Check:   9672  
Weight: 1.00
```

```
//
```

```
      seq1  agctagctag ctagct  
      seq2  aactaactaa ctaact
```