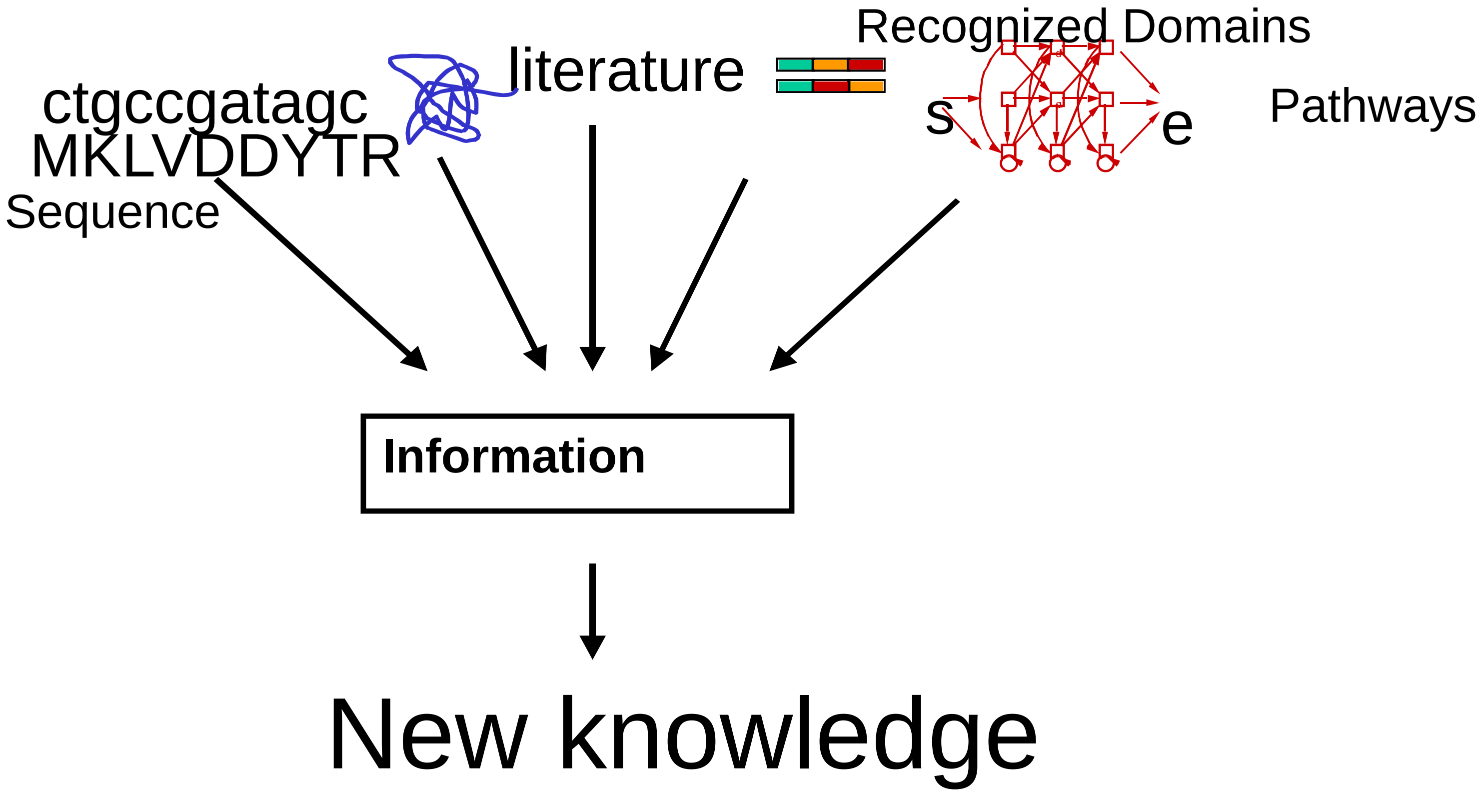
Basic Bioinformatic and Molecular Databases

Biological Data

Growth of biological informations from sequencing of genomes, coning of various genes, solution of protein structures has lead to establishment of specific databases

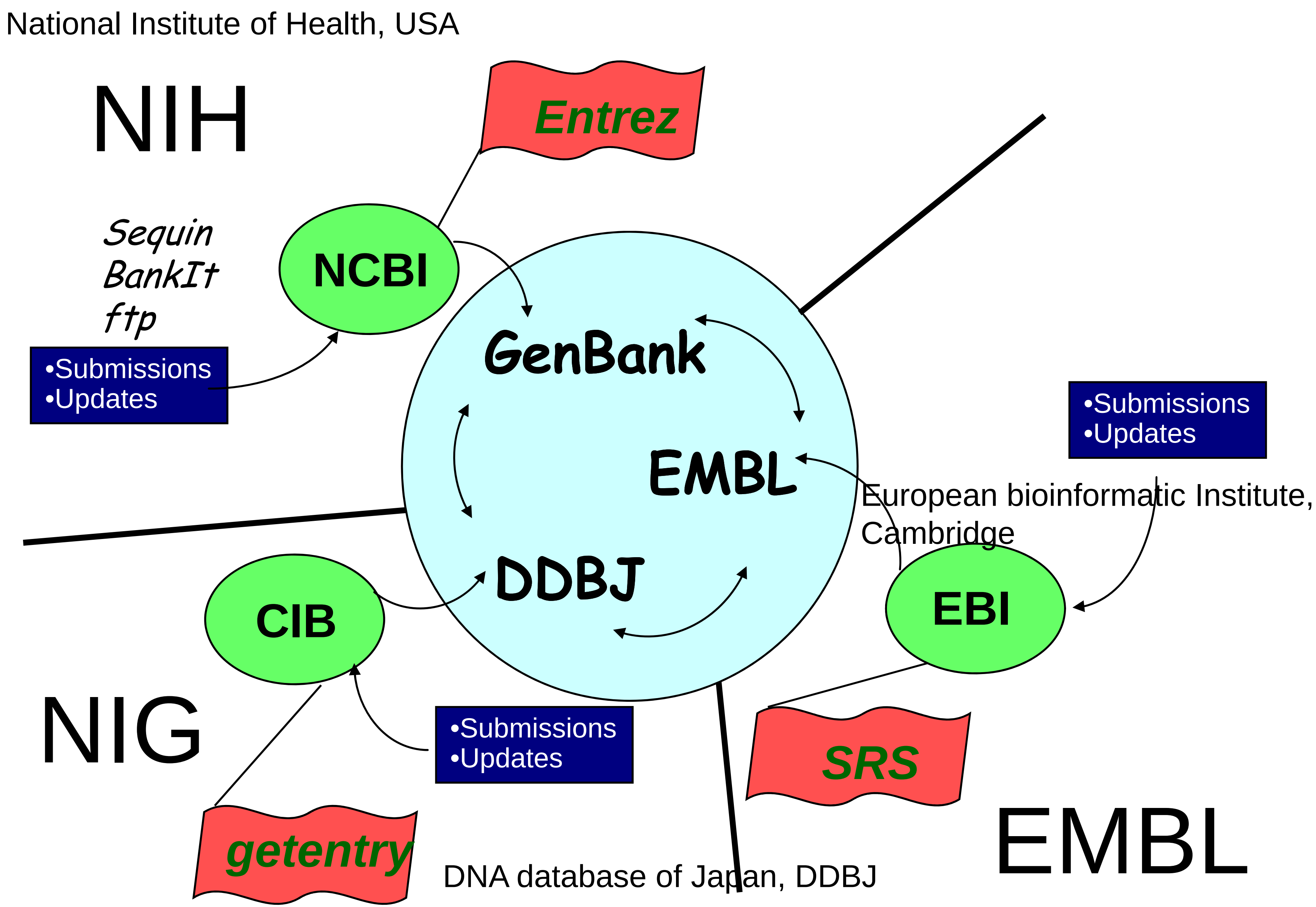


Databases

Based on source of data

- Primary database
- Secondary database
- Knowledge bases
- Integrated Database

Primary Databases



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 $x10^7$ sequences, over 3.2 $x10^{10}$ bp;

GenBank Format

NCBI Entrez Nucleotide

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display GenBank Save Text Add to Clipboard Get Subsequence

1: AF467571. Hydrobia acuta ac...[gi:22416476] Links

```
LOCUS       AF467571                 638 bp    DNA     linear   INV 22-AUG-2002
DEFINITION  Hydrobia acuta acuta isolate 1479 cytochrome c oxidase subunit I
            (COI) gene, partial cds; mitochondrial gene for mitochondrial
            product.
ACCESSION   AF467571
VERSION     AF467571.1  GI:22416476
KEYWORDS    .
SOURCE      Hydrobia acuta acuta.
  ORGANISM  Mitochondrion Hydrobia acuta acuta
            Eukaryota; Metazoa; Mollusca; Gastropoda; Caenogastropoda;
            Mesogastropoda; Rissooidea; Hydrobiidae; Hydrobia.
REFERENCE   1  (bases 1 to 638)
  AUTHORS   Wilke,T. and Pfenninger,M.
  TITLE     Separating historic events from recurrent processes in cryptic
            species: phylogeography of mud snails (Hydrobia spp.)
  JOURNAL   Mol. Ecol. 11 (8), 1439-1451 (2002)
   PUBMED   12144664
REFERENCE   2  (bases 1 to 638)
  AUTHORS   Wilke,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-JAN-2002) Department of Microbiology and Tropical
            Medicine, The George Washington University, 2300 Eye Street,
            Washington, DC 20037, USA
FEATURES             Location/Qualifiers
     source          1..638
                     /organism="Hydrobia acuta acuta"
                     /organelle="mitochondrion"
                     /isolate="1479"
                     /sub_species="acuta"
                     /db_xref="taxon:133416"
                     /country="Spain: Puerto de Mahon"
     gene            <1..>638
```

```
                     /translation="MGARGALLLALLLARAGLRKPESQEAAPLSGPCGRRVITSRIVG
                     GEDAELGRWPWQGSLRLWDSHVCGVSLLSHRWALTAAHCFETYSDLSDPSGWMVQFGQ
                     LTSMPSFWSLQAYYTRYFVSNIYLSPRYLGNSPYDIALVKLSAPVTYTKHIQPICLQA
                     STFEFENRTDCWVTGWGYIKEDEALPSPHTLQEVQVAIINNSMCNHLFLKYSFRKDIF
                     GDMVCAGNAQGGKDACFGDSGGPLACNKNGLWYQIGVVSWGVGCGRPNRPGVYTNISH
                     HFEWIQKLMAQSGMSQPDPSWPLLFFPLLWALPLLGPV"
     polyA site      1058
                     /note="13 a nucleotides"
BASE COUNT      208 a    317 c    305 g    241 t
ORIGIN
        1 gccatgggcg cgcgcggggc gctgctgctg gcgctgctgc tggctcgggc tggactcagg
       61 aagccggagt cgcaggaggc ggcgccgtta tcaggaccat gcggccgacg ggtcatcacg
      121 tcgcgcatcg tgggtggaga ggacgccgaa ctcgggcgtt ggccgtggca ggggagcctg
      181 cgcctgtggg attcccacgt atgcggagtg agcctgctca gccaccgctg ggcactcacg
      241 gcggcgcact gctttgaaac ctatagtgac cttagtgatc cctccgggtg gatggtccag
      301 tttggccagc tgacttccat gccatccttc tggagcctgc aggcctacta cacccgttac
      361 ttcgtatcga atatctatct gagccctcgc tacctgggga attcacccta tgacattgcc
      421 ttggtgaagc tgtctgcacc tgtcacctac actaaacaca tccagcccat ctgtctccag
```

GenBank format

NCBI Entrez Nucleotide

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxono

Search Nucleotide for [ ] Go Clear

Limits Preview/Index History Clipboard

Display FASTA Save Text Add to Clipboard Get Subsequence

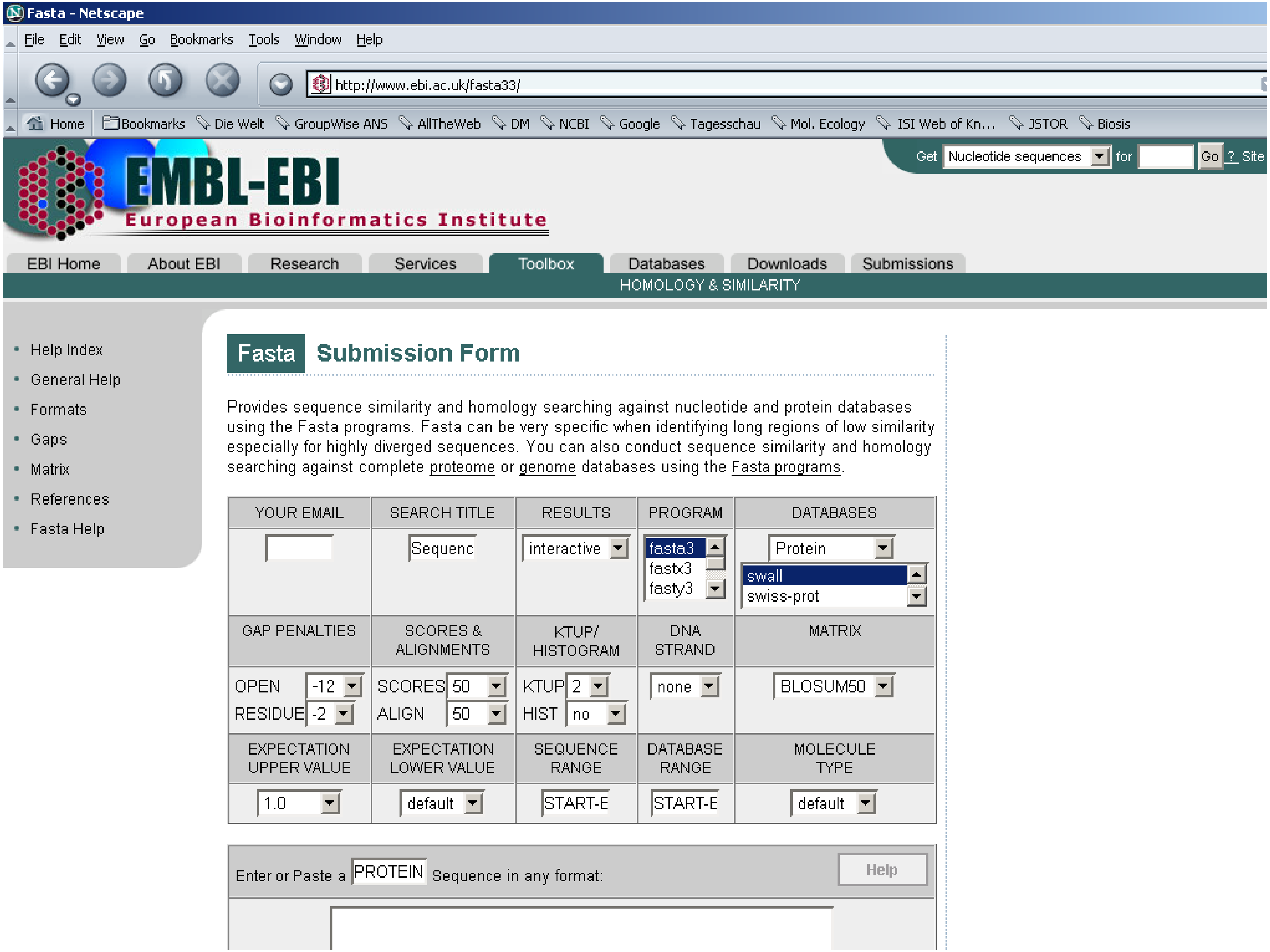
1: AF467571. Hydrobia acuta ac...[gi:22416476]

```
>gi|22416476|gb|AF467571.1| Hydrobia acuta acuta isolate 1479 cytochrome c oxidase subunit I (COI) gene, parti
ATTTTATTTGGTATGTGGTCTGGGTTAGTAGGTACAGCACTAAGTTTGTTAATTCGTGCTGAACTAGGTC
AGCCTGGTGCGCTTTTGGGTGATGATCAGCTTTATAACGTAATTGTTACTGCTCATGCCTTTGTTATAAT
TTTTTTTCTTGTAATGCCTATAATAATTGGTGGCTTTGGAAATTGATTAGTGCCTTTAATACTTGGTGCT
CCAGATATAGCTTTTCCTCGGCTTAATAACATAAGTTTCTGACTTTTACCTCCTGCTTTGCTATTATTAC
TTTCTTCGGCAGCTGTAGAGAGAGGAGCGGGGACAGGATGAACCGTGTATCCCCCATTATCTAGTAACAT
TGCTCACGCGGGGGGGTCTGTAGATTTAGCTATTTTTTCTCTCCACTTAGCGGGTGTTTCTTCTATTCTT
GGGGCTGTAAATTTTATTACAACTATCATTAATATACGGTGACGAGGAATGCAGTTTGAGCGGCTTCCGT
TGTTCGTATGATCTGTAAAAATTACTGCCATTCTATTATTACTATCTTTACCTGTCTTAGCTGGTGCTAT
TACTATGCTTTTAACGGATCGAAATTTTAATACTGCATTTTTCGACCCAGCAGGAGGTGGAGACCCTATT
TTATACCA
```

Revised: July 5, 2002.

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Fasta format


Fasta - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.ebi.ac.uk/fasta33/
Home Bookmarks Die Welt GroupWise ANS AllTheWeb DM NCBI Google Tagesschau Mol. Ecology ISI Web of Kn... JSTOR Biosis
Get Nucleotide sequences for Go ? Site
EMBL-EBI
European Bioinformatics Institute
EBI Home About EBI Research Services Toolbox Databases Downloads Submissions
HOMOLOGY & SIMILARITY
Help Index
General Help
Formats
Gaps
Matrix
References
Fasta Help
Fasta Submission Form
Provides sequence similarity and homology searching against nucleotide and protein databases using the Fasta programs. Fasta can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity and homology searching against complete proteome or genome databases using the Fasta programs.
YOUR EMAIL
SEARCH TITLE
Sequenc
RESULTS
interactive
PROGRAM
fasta3
fastx3
fasty3
DATABASES
Protein
swall
swiss-prot
GAP PENALTIES
OPEN -12
RESIDUE -2
SCORES & ALIGNMENTS
SCORES 50
ALIGN 50
KTUP/ HISTOGRAM
KTUP 2
HIST no
DNA STRAND
none
MATRIX
BLOSUM50
EXPECTATION UPPER VALUE
1.0
EXPECTATION LOWER VALUE
default
SEQUENCE RANGE
START-E
DATABASE RANGE
START-E
MOLECULE TYPE
default
Enter or Paste a PROTEIN Sequence in any format:
Help

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=MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLLE
FT                   AKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRG
FT                   QALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITAD
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
XX
SQ   Sequence 3398 BP; 698 A; 1034 C; 991 G; 675 T; 0 other;
     agcttctggg cttccagacc cagctacttt gcggaactca gcaacccagg catctctgag        60
     tctccgccca agaccgggat gccccccagg aggtgtccgg gagcccagcc tttcccagat       120
```

annotation

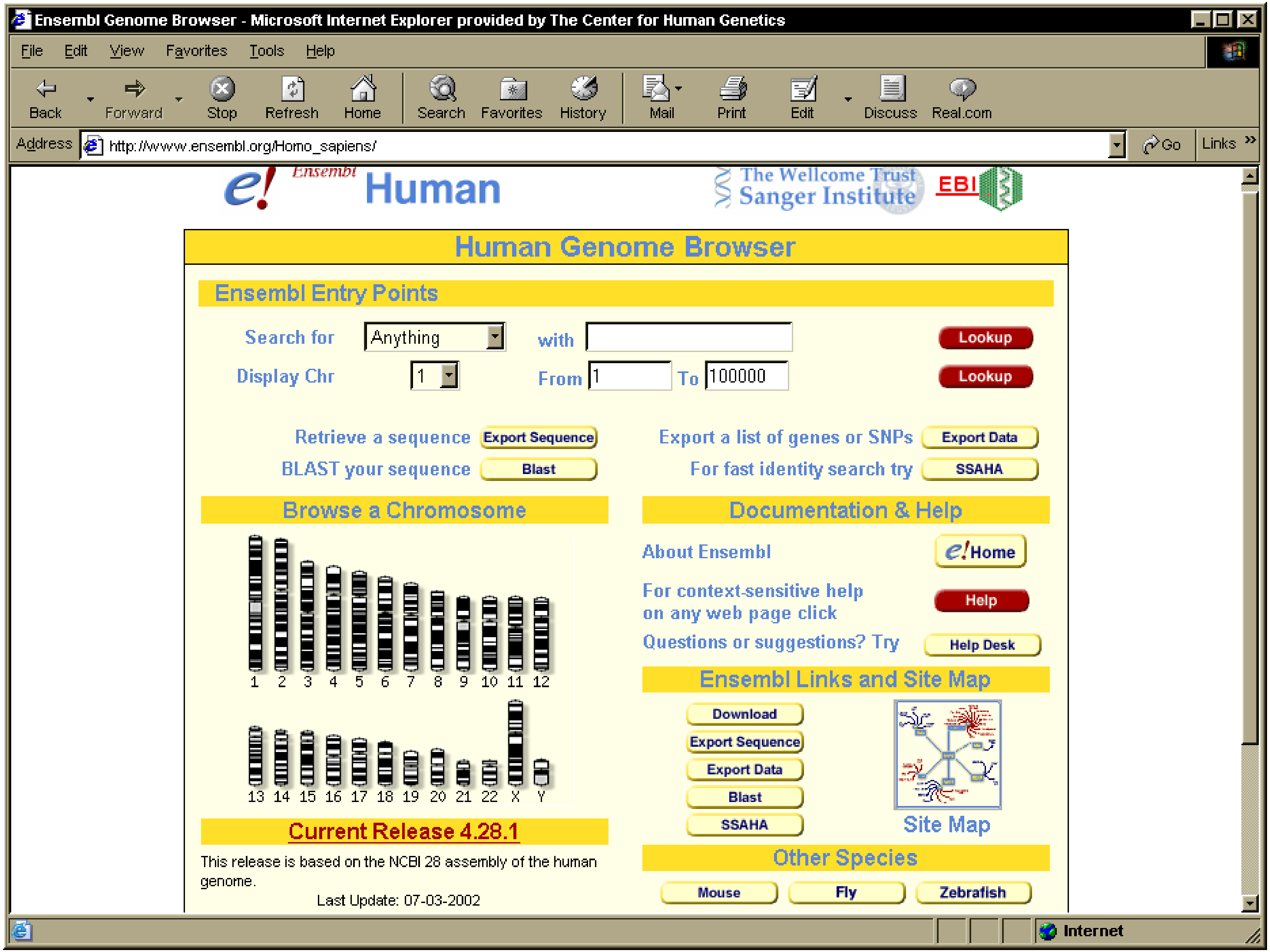
sequence

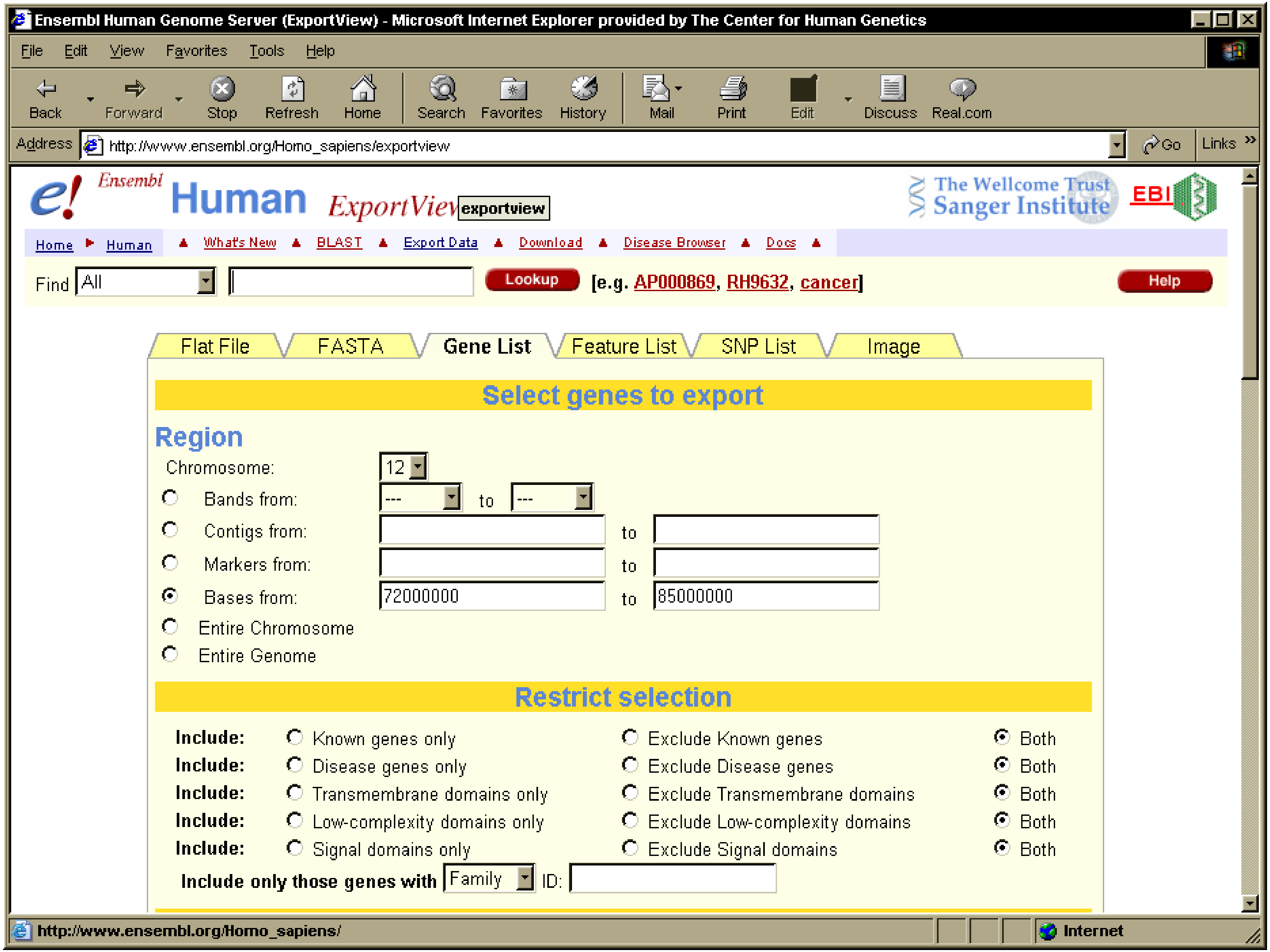
Databases related to Genomics

- 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).

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


Ensembl Genome Browser - Microsoft Internet Explorer provided by The Center for Human Genetics
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss Real.com
Address http://www.ensembl.org/Homo_sapiens/
Go Links
Ensembl
Human
The Wellcome Trust Sanger Institute
EBI
Human Genome Browser
Ensembl Entry Points
Search for Anything with Lookup
Display Chr 1 From 1 To 100000 Lookup
Retrieve a sequence Export Sequence
BLAST your sequence Blast
Export a list of genes or SNPs Export Data
For fast identity search try SSAHA
Browse a Chromosome
1 2 3 4 5 6 7 8 9 10 11 12
13 14 15 16 17 18 19 20 21 22 X Y
Current Release 4.28.1
This release is based on the NCBI 28 assembly of the human genome.
Last Update: 07-03-2002
Documentation & Help
About Ensembl Home
For context-sensitive help on any web page click Help
Questions or suggestions? Try Help Desk
Ensembl Links and Site Map
Download
Export Sequence
Export Data
Blast
SSAHA
Site Map
Other Species
Mouse Fly Zebrafish
Internet


Ensembl Human Genome Server (ExportView) - Microsoft Internet Explorer provided by The Center for Human Genetics
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss Real.com
Address http://www.ensembl.org/Homo_sapiens/exportview
Go Links
Ensembl Human ExportView exportview
The Wellcome Trust Sanger Institute
EBI
Home Human What's New BLAST Export Data Download Disease Browser Docs
Find All Lookup [e.g. AP000869, RH9632, cancer] Help
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:
http://www.ensembl.org/Homo_sapiens/
Internet

Protein Databases

Types of Data

– Sequence

– Function

– Recognized Domains

– Structural Information & Models

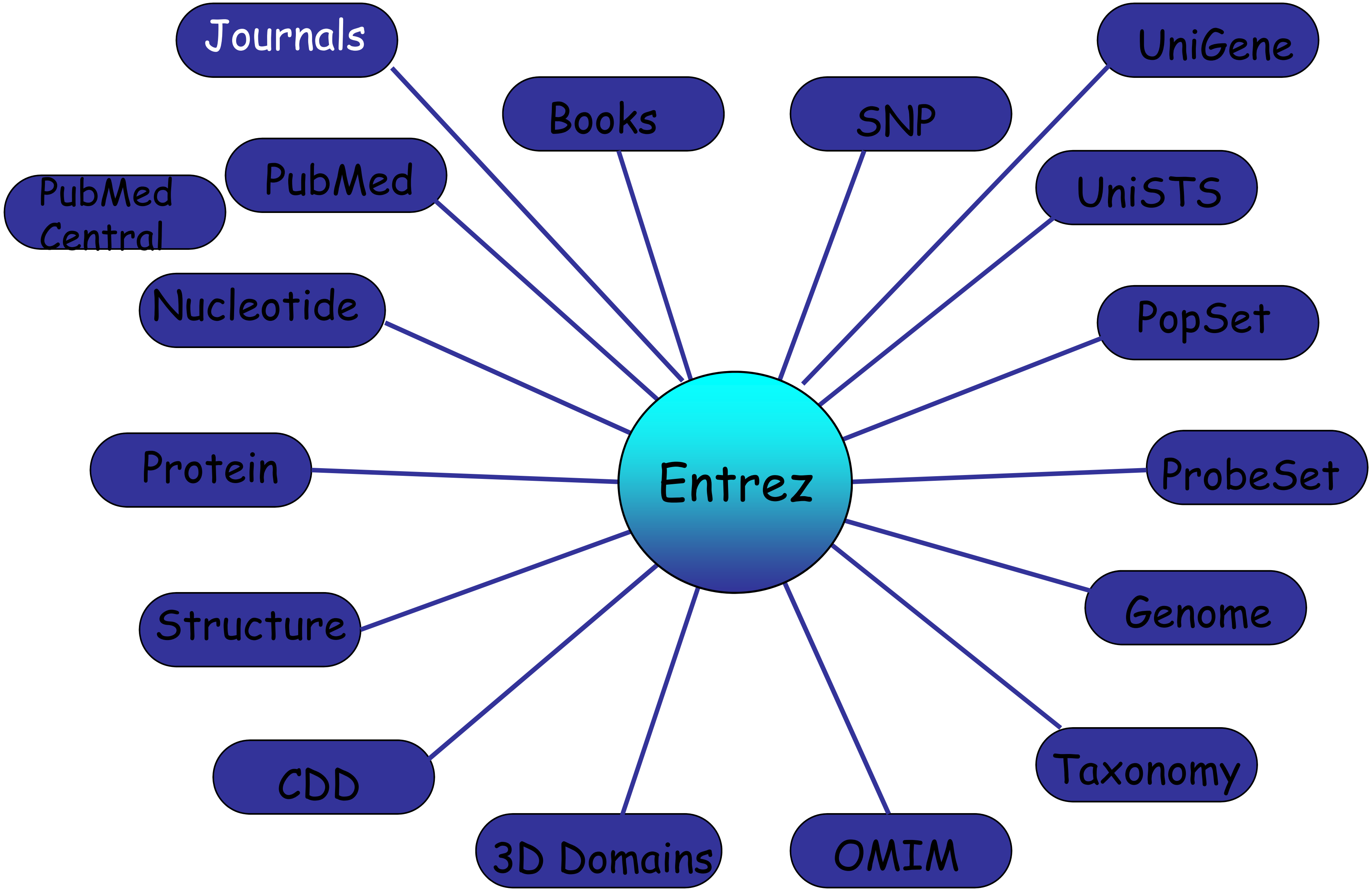
– Pathway Information & Models

Entrez-Protein

- Part of a much larger library
- Provided through the NIH
- Entrez provides Web and programmatic access

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein

The Expanding Entrez System



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

NCBI

National Library of Medicine PubMed

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy | OMIM

Search PubMed for HMM Go Clear

Limits | Preview/Index | History | Clipboard

About Entrez

Entrez PubMed

Overview
Help | FAQ
New/Noteworthy

PubMed Services
Journal Browser
MeSH Browser
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Cubby NEW

Related Resources
Order Documents
Grateful Med
Consumer Health
Clinical Alerts
ClinicalTrials.gov

Privacy Policy

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

PIR

- Protein Information Resource
- Arguably the most comprehensive

http://www-nbrf.georgetown.edu/pirwww/dbinfo/pirpsd.html

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 X 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

http://us.expasy.org/sprot/

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

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 YKAEEHTVVL TVTGEPCHFP FQYHRQLYHK

        70         80         90        100        110        120
         |          |          |          |          |          |
CTHKGRPGPQ PWCATTPNFD QDQRWGYCLE PKKVKDHCSK HSPCQKGGTC VNMPSGPHCL

       130        140        150        160        170        180
         |          |          |          |          |          |
CPQHLTGNHC QKEKCFEPQL LRFFHKNEIW YRTEQAAVAR CQCKGPDAHC QRLASQACRT
```

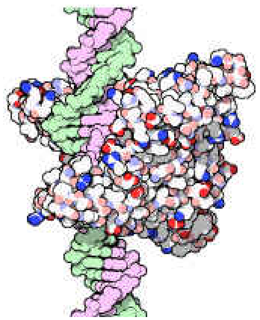
Structure Databases

- **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

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/

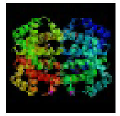


Effective use of PDB

- 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

PDB
PROTEIN DATA BANK

Structure Explorer - 4HHB



Summary Information

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 A 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	$C_{34}H_{32}N_4O_4FE_1$
PO4	PHOSPHATE ION	O_4P_1

Other Versions: **2HHB, 3HHB**

© RCSB

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: 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
………
```

BRENDA

- Comprehensive Enzyme Database

http://www.brenda.uni-koeln.de/

ENZYME

- Nomenclature database
- Hierarchical organization

http://us.expasy.org/enzyme/

KEGG

- Kyoto Encyclopedia of Genes and Genomes
- Its own classification scheme (KO)
- Massive pathway maps
- Links into COGS and ENZYME
- Has a SOAP-based programmatic interface

http://www.genome.ad.jp/kegg/

InterPro

- A large aggregated database, uses many sources in its search

http://www.ebi.ac.uk/interpro/

Secondary Structure Prediction

PSIPRED - a highly accurate method for protein secondary structure prediction MEMSAT and MEMSAT-SVM - widely used transmembrane topology prediction method and one of GenTHREADER, pGenTHREADER and pDomTHREADER - sequence profile based fold recognition methods.

http://bioinf.cs.ucl.ac.uk/psipred/

Helical Wheel Projections

http://rzlab.ucr.edu/scripts/wheel/wheel.cgi

PSI/Nature
Structural Biology Knowledge database

- Keep informed about advances in structural biology and structural genomics. Discover how protein sequences, three-dimensional structures and models relate to biological function. Stay up to date with the latest protocols, materials and technologies.
- Structure biology update
- Targets
- Protein structures and functions
- Membrane proteins
- Homology models
- Methods
- Publications

- http://sbkb.org/