Nucleotide and Protein sequence databases

Dinesh Gupta
Structural and Computational Biology Group
ICGEB
Nucleotide sequence databases

- EMBL, GenBank, and DDBJ are the three primary nucleotide sequence databases
  - EMBL [www.ebi.ac.uk/embl/](http://www.ebi.ac.uk/embl/)
  - DDBJ [www.ddbj.nig.ac.jp](http://www.ddbj.nig.ac.jp)
Genbank

• An annotated collection of all publicly available nucleotide and proteins

• Set up in 1979 at the LANL (Los Alamos).

• Maintained since 1992 NCBI (Bethesda).

Growth of the International Nucleotide Sequence Database Collaboration
EMBL Nucleotide Sequence Database

- An annotated collection of all publicly available nucleotide and protein sequences

- Created in 1980 at the *European Molecular Biology Laboratory* in Heidelberg.

- Maintained since 1994 by EBI- Cambridge.

- [http://www.ebi.ac.uk/embl.html](http://www.ebi.ac.uk/embl.html)
EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects, and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 102, Dec 2009), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in *Nucleic Acids Research 2009 37: D19-D25* provides further information and details.

The EMBL nucleotide sequence database forms part of the European Nucleotide Archive, an EBI project led by Guy Cochrane as part of the The Protein and Nucleotide Database Group (PANDA) under Ewan Birney.

<table>
<thead>
<tr>
<th>Link</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Database queries, Completed genomes webserver, FTP archives (EMBL release, alignments etc), EMBL sequence version archive (SVA), Browse by geography.</td>
</tr>
<tr>
<td>Submission</td>
<td>Primary sequence submissions, third party annotation, updates.</td>
</tr>
<tr>
<td>Publications</td>
<td>Group publications</td>
</tr>
<tr>
<td>People</td>
<td>Group members</td>
</tr>
<tr>
<td>Contact</td>
<td>How to contact the EMBL Nucleotide Sequence Database</td>
</tr>
<tr>
<td>News</td>
<td>List of recent changes on this site</td>
</tr>
</tbody>
</table>
This morning the EMBL Database contained 273,064,874,799 nucleotides in 174,853,084 entries.

Breakdown by entry type:

<table>
<thead>
<tr>
<th>Entry Type</th>
<th>Entries</th>
<th>Nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>114,466,528</td>
<td>110,775,380,736</td>
</tr>
<tr>
<td>Constructed (CON)</td>
<td>5,976,255</td>
<td>n/a</td>
</tr>
<tr>
<td>Third Party Annotation (TPA)</td>
<td>6,593</td>
<td>372,891,884</td>
</tr>
<tr>
<td>Whole Genome Shotgun (WGS)</td>
<td>54,403,706</td>
<td>152,516,602,179</td>
</tr>
</tbody>
</table>

Top Organisms

By nucleotide count

- Homo sapiens
- Mus musculus
- Rattus norvegicus
- Drosophila melanogaster
- C. elegans
- Saccharomyces cerevisiae
- Caenorhabditis elegans
- Danio rerio
- Other

By entry count

- Homo sapiens
- Mus musculus
- Rattus norvegicus
- Drosophila melanogaster
- C. elegans
- Saccharomyces cerevisiae
- Caenorhabditis elegans
- Danio rerio
- Other
DDBJ-DNA Data Bank of Japan

• An annotated collection of all publicly available nucleotide and protein sequences

• Started, 1984 at the National Institute of Genetics (NIG) in Mishima.

• Still maintained in this institute a team led by Takashi Gojobori.

• [http://www.ddbj.nig.ac.jp](http://www.ddbj.nig.ac.jp)
DDBJ: DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of the three summit databanks that construct DDBJ/EMBL/GenBank International Nucleotide Sequence Database, which was established through cooperative work with EBI in Europe and NCBI in USA.

Photo by Tatsuko Kawasaki

Hot Topics

- Jan. 15, 2010: DAD (DDBJ amino acid database) Rel. 50.0 Released
- Jan. 7, 2010: Release of new false killer whale (Pseudorca crassidens) GSS 90,007 entries
- Jan. 4, 2010: DDBJ/EMBL/GenBank Feature Table Definition revised

Maintenance

- Jan. 15, 2009: ARSA database search (DDBJ, DAD) temporary unavailable
- Dec. 28, 2009: Apology for the failure of ClustalW running
- Dec. 8, 2009: Closing H-Inv DB mirror site

Sequence Data Submission

- Submit my sequences
  Orientation for the data submission
- Update my entries
  Guidance for the update of the entry

FTP/Web API

- FTP (ftp.ddbj.nig.ac.jp)
  Download data files
- Web API
  Programmatic interfaces of DDBJ Web services

DNA Data Bank of Japan (DDBJ)

- Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ)
Number of entries
(current 174,853,084)

Graphs created on 15 January 2010
Sequence submission

• Data mainly direct submissions from the authors.

• Submissions through the Internet:
  – Web forms.
  – Email.

• Sequences shared/exchanged between the 3 centers on a daily basis:
  – The sequence content of the banks is identical.
Derived databases

- CUTG Codon usage tabulated from GenBank
  http://www.kazusa.or.jp/codon/
- Genetic Codes Deviations from the standard genetic code in various organisms and organelles
- TIGR Gene Indices Organism-specific databases of EST and gene sequences
  http://www.tigr.org/tdb/tgi.shtml
- UniGene Unified clusters of ESTs and full-length mRNA sequences
- ASAP Alternative spliced isoforms
  http://www.bioinformatics.ucla.edu/ASAP
- Intronerator Introns and alternative splicing in C.elegans and C.briggsae
  http://www.cse.ucsc.edu/~kent/intronerator/
Sequence Retrieval Tools

• Various tools to get sequences of interests from databases
  – SRS for EMBL and other DBs [http://srs.ebi.ac.uk](http://srs.ebi.ac.uk)
  – Fetch in GCG package
  – Seqret in EMBOSS
Protein Databases

• General Sequence databases
• Protein properties
• Protein localization and targeting
• Protein sequence motifs and active sites
• Protein domain databases; protein classification
• Databases of individual protein families
Protein Databases

• General Sequence databases
• Protein properties
• Protein localization and targeting
• Protein sequence motifs and active sites
• Protein domain databases; protein classification
• Databases of individual protein families
NCBI Protein database

• The NCBI Entrez Protein database Sequences from: SwissProt, the Protein Information Resource, the Protein Research Foundation, the Protein Data Bank, and translations from annotated coding regions in the GenBank and RefSeq databases.

• Protein sequence records in Entrez have links to pre-computed protein BLAST alignments, protein structures, conserved protein domains, nucleotide sequences, genomes, and genes.

Swiss-Prot

- The Swiss-Prot Protein Knowledgebase is a curated protein sequence database established in 1986.
- It provides a high level of annotation (such as the description of protein function, domains structure, post-translational modifications, variants, etc.), a **minimal level of redundancy** and high level of integration with other databases.
- Now, it is part of the Universal Protein Knowledgebase (a part of UniProt), a "one-stop shop" that allows easy access to all publicly available information of protein sequence annotation.
The Swiss-Prot, TrEMBL, and PIR protein database activities have united to form the Universal Protein Resource (UniProt)

- **Uniprot Knowledgebase (UniprotKB):** curated Sequence information, annotations, linked to other databases.

- **Uniprot Reference Clusters (UniRef):** removing sequence redundancy by merging sequences that are 100%, 90% and 50%, no annotations, linked to Knowledgebase and UniParc records.

- **Uniprot Archive (UniParc):** history of sequences, no annotation, linked to source records.
Welcome

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

<table>
<thead>
<tr>
<th>UniProtKB</th>
<th>Protein knowledgebase, consists of two sections:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>⭐ Swiss-Prot, which is manually annotated and reviewed.</td>
</tr>
<tr>
<td></td>
<td>⭐ TrEMBL, which is automatically annotated and is not reviewed.</td>
</tr>
<tr>
<td></td>
<td>Includes Complete Proteome Sets.</td>
</tr>
<tr>
<td>UniRef</td>
<td>Sequence clusters, used to speed up similarity searches.</td>
</tr>
<tr>
<td>UniParc</td>
<td>Sequence archive, used to keep track of sequences and their identifiers.</td>
</tr>
<tr>
<td>Supporting data</td>
<td>Literature citations, taxonomy, keywords and more.</td>
</tr>
</tbody>
</table>
UniprotKB/Swiss-Prot protein knowledgebase release 57.12 statistics

1. INTRODUCTION

Release 57.12 of 15-Dec-09 of UniprotKB/Swiss-Prot contains 513877 sequence entries, comprising 180750753 amino acids abstracted from 185524 references.

897 sequences have been added since release 57.11, the sequence data of 59 existing entries has been updated and the annotations of 429352 entries have been revised.

Number of fragments: 8437
Number of additional sequences produced by alternative splicing, initiation or promoter usage, or ribosomal frameshifting: 28670

Protein existence (PE):

<table>
<thead>
<tr>
<th></th>
<th>entries</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>Evidence at protein level</td>
<td>67947</td>
</tr>
<tr>
<td>2:</td>
<td>Evidence at transcript level</td>
<td>66340</td>
</tr>
<tr>
<td>3:</td>
<td>Inferred from homology</td>
<td>363688</td>
</tr>
<tr>
<td>4:</td>
<td>Predicted</td>
<td>14332</td>
</tr>
<tr>
<td>5:</td>
<td>Uncertain</td>
<td>1570</td>
</tr>
</tbody>
</table>

The growth of the database is summarized below.
2.3 Taxonomic distribution of the sequences

![Pie chart showing distribution of sequences by kingdom.]

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>sequences (% of the database)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea</td>
<td>18172 (4%)</td>
</tr>
<tr>
<td>Bacteria</td>
<td>322755 (63%)</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>158145 (31%)</td>
</tr>
<tr>
<td>Viruses</td>
<td>14805 (3%)</td>
</tr>
</tbody>
</table>

Within Eukaryota:

![Pie chart showing distribution of sequences within Eukaryota.]

<table>
<thead>
<tr>
<th>Category</th>
<th>sequences (% of Eukaryota) (% of the complete database)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fungi</td>
<td>(16%)</td>
</tr>
<tr>
<td>Viridiplanta</td>
<td>(18%)</td>
</tr>
<tr>
<td>Other</td>
<td>(8%)</td>
</tr>
<tr>
<td>Insects</td>
<td>(5%)</td>
</tr>
<tr>
<td>Nematoda</td>
<td>(2%)</td>
</tr>
<tr>
<td>Other Vertebra</td>
<td>(10%)</td>
</tr>
<tr>
<td>Homo</td>
<td>(15%)</td>
</tr>
<tr>
<td>Other Mammals</td>
<td>(28%)</td>
</tr>
</tbody>
</table>
The average sequence length in UniProtKB/Swiss-Prot is 351 amino acids.
Amino acid composition

Legend: gray = aliphatic, red = acidic, green = small hydroxy, blue = basic, black = aromatic, white = amide, yellow = sulfur
• The shortest sequence is GWA_SEPOF (P83570): 2 amino acids, a Neuropeptide from cuttle fish.

• The longest sequence is TITIN_MOUSE (A2ASS6): 35213 amino acids, assembly and functioning of vertebrate striated muscles, defects cause myopathies.
http://www.expasy.org/sprot

Swiss-Prot
Protein knowledgebase
TrEMBL
Computer-annotated supplement to Swiss-Prot

The UniProt Knowledgebase consists of:
- UniProtKB/Swiss-Prot: a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases (More details / References / Linking to UniProtKB/Swiss-Prot / User manual / Recent changes / Disclaimer).
- UniProtKB/TrEMBL: a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Knowledgebase Release 15.12 consists of:
UniProtKB/Swiss-Prot Release 57.12 of 15-Dec-09: 513877 entries (More statistics)
UniProtKB/TrEMBL Release 40.12 of 15-Dec-09: 10034121 entries (More statistics)

Access to the UniProt Knowledgebase
- UniProt web site
- UniProtKB Taxonomy browser
- ViralZone - Portal to viral UniProtKB/Swiss-Prot entries
- BLAST similarity search
WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

<table>
<thead>
<tr>
<th>UniProtKB</th>
<th>Protein knowledgebase, consists of two sections:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>💫 Swiss-Prot, which is manually annotated and reviewed.</td>
</tr>
<tr>
<td></td>
<td>💫 TriEMBL, which is automatically annotated and is not reviewed.</td>
</tr>
<tr>
<td></td>
<td>Includes Complete Proteome Sets.</td>
</tr>
<tr>
<td>UniRef</td>
<td>Sequence clusters, used to speed up similarity searches.</td>
</tr>
<tr>
<td>UniParc</td>
<td>Sequence archive, used to keep track of sequences and their identifiers.</td>
</tr>
<tr>
<td>Supporting data</td>
<td>Literature citations, taxonomy, keywords and more.</td>
</tr>
</tbody>
</table>
**Accession** | **Entry name** | **Status** | **Protein names** | **Gene names** | **Organism** | **Length**
---|---|---|---|---|---|---
Q9YBY0 | NS1BP_HUMAN | | Influenza virus NS1A-binding protein | IIVNS1ABP (ARA3) (FLA3A) (KIAA0850) (NS1) (NS1BP) (HSPC068) | Homo sapiens (Human) | 642
P09222 | MX1_MOUSE | | Interferon-induced GTP binding protein Mx1 | Mx1 | Mus musculus (Mouse) | 631
Q920Q8 | NS1BP_MOUSE | | Influenza virus NS1A-binding protein homolog | IIVNS1abp (Kiasa0850) (Nd1) (Nd1L) (Nd1S) (Ns1) (Ns1bp) | Mus musculus (Mouse) | 642
Q5RG22 | NS1BA_DANRE | | Influenza virus NS1A-binding protein homolog ... | iivns1abpa (sickeyp-11g8.1) | Danio rerio (Zebrafish) (Brachydanio rerio) | 643
Q7VQ07 | NS1BB_DANRE | | Influenza virus NS1A-binding protein homolog ... | ivns1abpb | Danio rerio (Zebrafish) (Brachydanio rerio) | 640
Q6DFU2 | NS1BP_XENLA | | Influenza virus NS1A-binding protein homolog | ivns1abp | Xenopus laevis (African clawed frog) | 638
Q499Q3 | Q499Q3_RAT | | Myxovirus (Influenza virus) resistance 2 | Mx2 | Rattus norvegicus (Rat) | 659
Q499S4 | Q499S4_RAT | | Myxovirus (Influenza virus) resistance 1 | Mx1 | Rattus norvegicus (Rat) | 652
B8JK03 | B8JK03_DANRE | | Myxovirus (Influenza) resistance A | mxa (DKEY-197A6.3-001) | Danio rerio (Zebrafish) (Brachydanio rerio) | 632
Q1ECX1 | Q1ECX1_DANRE | | Myxovirus (Influenza) resistance A | mxa (zgc:136725) | Danio rerio (Zebrafish) (Brachydanio rerio) | 632
<table>
<thead>
<tr>
<th>Accession</th>
<th>Entry name</th>
<th>Status</th>
<th>Protein names</th>
<th>Gene names</th>
<th>Organism</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q9Y6Y0</td>
<td>NS1BP_HUMAN</td>
<td></td>
<td>Influenza virus NS1A-binding protein</td>
<td>IVNS1ABP (ARA3) (FLARA3) (KIAA0850) (NS1) (NS1BP) (HSPC686)</td>
<td>Homo sapiens (Human)</td>
<td>642</td>
</tr>
<tr>
<td>P09522</td>
<td>MX1_MOUSE</td>
<td></td>
<td>Interferon-induced GTP-binding protein Mx1</td>
<td>Mx1</td>
<td>Mus musculus (Mouse)</td>
<td>631</td>
</tr>
<tr>
<td>Q926G8</td>
<td>NS1BP_MOUSE</td>
<td></td>
<td>Influenza virus NS1A-binding protein homolog</td>
<td>Ivns1ab (钾as0850) (Nd1) (Nd1L) (Nd1S) (Ns1) (Ns1bp)</td>
<td>Mus musculus (Mouse)</td>
<td>642</td>
</tr>
<tr>
<td>Q5R82</td>
<td>NS10A_DANRE</td>
<td></td>
<td>Influenza virus NS1A-binding protein homolog ...</td>
<td>Ivns1abpa (si dkeyp-11g8.1)</td>
<td>Danio rerio (Zebrafish) (Brachydanio rerio)</td>
<td>643</td>
</tr>
<tr>
<td>Q7VQ8</td>
<td>NS11B_DANRE</td>
<td></td>
<td>Influenza virus NS1A-binding protein homolog ...</td>
<td>Ivns1abpb</td>
<td>Danio rerio (Zebrafish) (Brachydanio rerio)</td>
<td>640</td>
</tr>
<tr>
<td>Q6DFU2</td>
<td>NS1BP_XENLA</td>
<td></td>
<td>Influenza virus NS1A-binding protein homolog</td>
<td>Ivns1ab</td>
<td>Xenopus laevis (African clawed frog)</td>
<td>638</td>
</tr>
<tr>
<td>P03485</td>
<td>M1_A3A1</td>
<td></td>
<td>Matrix protein 1</td>
<td>M</td>
<td>Influenza A virus (strain A/Puerto Rico/8/1934 H1N1)</td>
<td>252</td>
</tr>
<tr>
<td>P03437</td>
<td>HEMX_68A0</td>
<td></td>
<td>Hemagglutinin</td>
<td>HA</td>
<td>Influenza A virus (strain A/AmCh/2/1968 H3N2)</td>
<td>566</td>
</tr>
<tr>
<td>P03495</td>
<td>NS1_172A2</td>
<td></td>
<td>Non-structural protein 1</td>
<td>NS</td>
<td>Influenza A virus (strain A/AmOr/307/1972 H3N2)</td>
<td>237</td>
</tr>
<tr>
<td>P06620</td>
<td>NRAM_67A0</td>
<td></td>
<td>Neuraminidase</td>
<td>NA</td>
<td>Influenza A virus (strain A/Tokyo/3/1967 H2N2)</td>
<td>469</td>
</tr>
<tr>
<td>P03472</td>
<td>NRAM_175A5</td>
<td></td>
<td>Neuraminidase</td>
<td>NA</td>
<td>Influenza A virus (strain A/Tam/Australia/G70/C/1975 H11N9)</td>
<td>470</td>
</tr>
<tr>
<td>P05777</td>
<td>M1_33A0</td>
<td></td>
<td>Matrix protein 1</td>
<td>M</td>
<td>Influenza A virus (strain A/Wilson-Smith/1933 H1N1) (Influenza A virus (strain A/Ws/1933 H1N1))</td>
<td>252</td>
</tr>
<tr>
<td>Accession</td>
<td>Entry Name</td>
<td>Status</td>
<td>Protein Names</td>
<td>Genes</td>
<td>Organism</td>
<td>Length</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>--------</td>
<td>--------------------------------------------------------------------------------</td>
<td>----------------------------</td>
<td>-----------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Q9Y6Y0</td>
<td>NS1BP_HUMAN</td>
<td>⭐</td>
<td>Influenza virus NS1A-binding protein (NS1-binding protein) (NS1-BP) (Aryl hydrocarbon receptor-associated protein 3)</td>
<td>IVNS1ABP (ARA3) (FLARA3) (KIAA0850) (NS1) (NS1BP) (HSPC068)</td>
<td>Homo sapiens (Human)</td>
<td>642</td>
</tr>
<tr>
<td>P03437</td>
<td>HEMA_68A0</td>
<td>⭐</td>
<td>Hemagglutinin precursor [Cleaved into Hemagglutinin HA1 chain; Hemagglutinin HA2 chain]</td>
<td>HA</td>
<td>Influenza A virus (strain A/Hong Kong/68 H3N2)</td>
<td>566</td>
</tr>
<tr>
<td>P03485</td>
<td>M1_J34A1</td>
<td>⭐</td>
<td>Matrix protein 1 (M1)</td>
<td>M</td>
<td>Influenza A virus (strain A/Hong Kong/68 H3N2)</td>
<td>252</td>
</tr>
</tbody>
</table>
1 - 25 of 91 results for **Influenza** AND keyword:**"3D-structure [2]** in UniProtKB sorted by **score** descending

- **Restrict term** "influenza" to **author**, **protein family**, **protein name**, **organism**, **strain**, **taxonomy**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Entry Name</th>
<th>Status</th>
<th>Protein Names</th>
<th>Genes</th>
<th>Organism</th>
<th>Length</th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>P03437</td>
<td>HEMA_I68A0</td>
<td>⭐</td>
<td>Hemagglutinin precursor</td>
<td>HA</td>
<td>Influenza A virus (strain A/Aichi/2/1968 H3N2)</td>
<td>586</td>
<td>X-ray crystallography (22)</td>
</tr>
<tr>
<td>P03485</td>
<td>M1_I34A1</td>
<td>⭐</td>
<td>Matrix protein 1 (M1)</td>
<td>M</td>
<td>Influenza A virus (strain A/Puerto Rico/8/1934 H1N1)</td>
<td>252</td>
<td>X-ray crystallography (3)</td>
</tr>
<tr>
<td>P06820</td>
<td>NRAM_I67A0</td>
<td>⭐</td>
<td>Neuraminidase (EC 3.2.1.18)</td>
<td>NA</td>
<td>Influenza A virus (strain A/Tokyo/3/1967 H2N2)</td>
<td>469</td>
<td>X-ray crystallography (11)</td>
</tr>
<tr>
<td>P03472</td>
<td>NRAM_I75A5</td>
<td>⭐</td>
<td>Neuraminidase (EC 3.2.1.18)</td>
<td>NA</td>
<td>Influenza A virus (strain A/Term/Australia/G70C/1975 H11N9)</td>
<td>470</td>
<td>X-ray crystallography (38)</td>
</tr>
<tr>
<td>P03495</td>
<td>NS1_I72A2</td>
<td>⭐</td>
<td>Non-structural protein 1 (NS1) (NS1A)</td>
<td>NS</td>
<td>Influenza A virus (strain A/Udom/307/1972 H3N2)</td>
<td>237</td>
<td>NMR spectroscopy (1) X-ray crystallography (1)</td>
</tr>
<tr>
<td>P63231</td>
<td>M2_I72A2</td>
<td>⭐</td>
<td>Matrix protein 2 (Proton channel protein M2)</td>
<td>M</td>
<td>Influenza A virus (strain A/Udom/307/1972 H3N2)</td>
<td>97</td>
<td>NMR spectroscopy (1)</td>
</tr>
</tbody>
</table>
### Names and origin

#### Protein names

- **Hemagglutinin [Precursor]**
  - Cleaved into:
    - Hemagglutinin HA1 chain
    - Hemagglutinin HA2 chain

#### Gene names

- **Name:** HA

#### Organism

- **Influenza A virus (strain A/Anchorage/2/1968 H3N2)**

#### Taxonomic identifier

- 387139 [NCBI]

#### Taxonomic lineage

- Viruses $\rightarrow$ ssRNA negative-strand viruses $\rightarrow$ Orthomyxoviridae $\rightarrow$ Influenzavirus A

#### Virus host

- Aves [TaxID: 8782]
- Homo sapiens (Human) [TaxID: 9606]
- Phocidae (true seals) [TaxID: 9709]
- Celacea (whales) [TaxID: 9721]
- Sus scrofa (Pig) [TaxID: 9623]

#### Protein existence

- Evidence at protein level.
### General annotation (Comments)

**Function**
Binds to sialic acid-containing receptors on the cell surface, bringing about the attachment of the virus particle to the cell. Plays a major role in the determination of host range restriction and virulence. Class I viral fusion protein. Responsible for penetration of the virus into the cell cytoplasm by mediating the fusion of the membrane of the endocytosed virus particle with the endosomal membrane. Low pH in endosomes induce an irreversible conformational change in HA2, releasing the fusion hydrophobic peptide. Several trimers are required to form a competent fusion pore.

**Subunit structure**
Homotrimer of disulfide-linked HA1-HA2.

**Subcellular location**

**Post-translational modification**
- In natural infection, inactive HA is matured into HA1 and HA2 outside the cell by one or more trypsin-like, arginine-specific endoprotease secreted by the bronchial epithelial cells. One identified protease that may be involved in this process is secreted in lungs by Clara cells. (Palmitylated [By similarity])

**Miscellaneous**
- Major glycoprotein, comprises over 80% of the envelope proteins present in virus particle.
- The extent of infection into host organism is determined by HA. Influenza viruses bud from the apical surface of polarized epithelial cells (e.g. bronchial epithelial cells) into lumen of lungs and are therefore usually pneumotropic. The reason is that HA is cleaved by trypsin Clara which is restricted to lungs. However, HAs of H5 and H7 pantropic avian viruses subtypes can be cleaved by furrin and subtilisin-type enzymes, allowing the virus to grow in other organs than lungs.
- The influenza A genome consist of 8 RNA segments. Genetic variation of hemagglutinin and/or neuraminidase genes results in the emergence of new influenza strains. The mechanism of variation can be the result of point mutations or the result of genetic reassortment between segments of two different strains.

### Sequence similarities
Belongs to the influenza viruses hemagglutinin family.

---

### Ontologies

#### Keywords

**Biological process**
- Fusion protein

**Cellular component**
- Envelope protein
- Membrane
### Molecule processing

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal peptide</td>
<td>1 - 16</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chain</td>
<td>17 - 344</td>
<td>328</td>
<td>Hemagglutinin HA1 chain</td>
<td></td>
</tr>
<tr>
<td>Chain</td>
<td>346 - 566</td>
<td>221</td>
<td>Hemagglutinin HA2 chain</td>
<td></td>
</tr>
</tbody>
</table>

### Regions

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topological domain</td>
<td>17 - 530</td>
<td>514</td>
<td>Extracellular</td>
<td></td>
</tr>
<tr>
<td>Transmembrane</td>
<td>531 - 551</td>
<td>21</td>
<td>(Potential)</td>
<td></td>
</tr>
<tr>
<td>Topological domain</td>
<td>552 - 566</td>
<td>15</td>
<td>Cytoplasmic (Potential)</td>
<td></td>
</tr>
</tbody>
</table>

### Sites

<table>
<thead>
<tr>
<th>Site</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>345 - 346</td>
<td>2</td>
<td>Cleavage, by host</td>
<td></td>
</tr>
</tbody>
</table>

### Amino acid modifications

<table>
<thead>
<tr>
<th>Modification</th>
<th>Position</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipidation</td>
<td>555</td>
<td>1</td>
<td>S-palmitoyl cysteine; by host</td>
<td></td>
</tr>
<tr>
<td>Lipidation</td>
<td>562</td>
<td>1</td>
<td>S-palmitoyl cysteine; by host (By similarity)</td>
<td></td>
</tr>
<tr>
<td>Lipidation</td>
<td>565</td>
<td>1</td>
<td>S-palmitoyl cysteine; by host (By similarity)</td>
<td></td>
</tr>
<tr>
<td>Glycosylation</td>
<td>24</td>
<td>1</td>
<td>N-linked (GlcNAc...)</td>
<td></td>
</tr>
<tr>
<td>Glycosylation</td>
<td>38</td>
<td>1</td>
<td>N-linked (GlcNAc...)</td>
<td></td>
</tr>
<tr>
<td>Glycosylation</td>
<td>54</td>
<td>1</td>
<td>N-linked (GlcNAc...)</td>
<td></td>
</tr>
<tr>
<td>Glycosylation</td>
<td>97</td>
<td>1</td>
<td>N-linked (GlcNAc...)</td>
<td></td>
</tr>
<tr>
<td>Glycosylation</td>
<td>181</td>
<td>1</td>
<td>N-linked (GlcNAc...)</td>
<td></td>
</tr>
<tr>
<td>Glycosylation</td>
<td>301</td>
<td>1</td>
<td>N-linked (GlcNAc...)</td>
<td></td>
</tr>
</tbody>
</table>
# General protein sequence databases

<table>
<thead>
<tr>
<th>Protein Sequence database</th>
<th>Source</th>
<th>Properties worth mentioning</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXProt</td>
<td>Proteins with experimentally verified function</td>
<td>Non redundant</td>
<td><a href="http://www.cmbi.kun.nl/EXProt/">http://www.cmbi.kun.nl/EXProt/</a></td>
</tr>
<tr>
<td>MIPS</td>
<td>Proteins from genome sequencing projects</td>
<td>Manually curated</td>
<td><a href="http://mips.gsf.de/">http://mips.gsf.de/</a></td>
</tr>
<tr>
<td>PIR</td>
<td>multiple source</td>
<td>Annotated sequences</td>
<td>merged with Uniprot now</td>
</tr>
<tr>
<td>PRF</td>
<td>sequences, source includes literature search</td>
<td>includes sequences not found in EMBL, Genbank and SwissProt also includes synthetic proteins and peptides</td>
<td><a href="http://www.prf.or.jp/en">http://www.prf.or.jp/en</a></td>
</tr>
<tr>
<td>Swiss-Prot</td>
<td>multiple</td>
<td>High level annotation and minimal level of redundancy</td>
<td><a href="http://www.expasy.org/sprot">http://www.expasy.org/sprot</a></td>
</tr>
</tbody>
</table>
Protein Databases

- General Sequence databases
- Protein properties
- Protein localization and targeting
- Protein sequence motifs and active sites
- Protein domain databases; protein classification
- Databases of individual protein families
DBs based on Protein properties

- AAindex: AAindex is a database of amino acid indices and amino acid mutation matrices
- Cybase: Cyclic proteins
- dbPTM: protein post-translational modification (PTM) information
- iProLINK: Integrated Protein Literature, INformation and Knowledge
- PFD - Protein Folding Database
- PINT: Protein-protein Interactions Thermodynamic Database
- PPD: Protein pKa Database
- ProTherm: Thermodynamic database for proteins and mutants
- REFOLD: Data related to refolding experiments
Protein Databases

• General Sequence databases
• Protein properties
• **Protein localization and targeting**
• Protein sequence motifs and active sites
• Protein domain databases; protein classification
• Databases of individual protein families
Protein localization and targeting

- DBSubLoc - Database of protein Subcellular Localization
- LOCATE: manually curated, immunofluorescence-based assay data
- MitoNuc: database of nuclear encoded mitochondrial proteins in Metazoa
- NESbase: Leucine-rich nuclear export signal
- NLSdb: database of nuclear localization signals
- NMPdb - Nuclear matrix associated proteins database
- NOPdb - Nucleolar Proteome Database:
- NPD - Nuclear Protein Database: results from MS
- Nuclear Receptor Resource:
  - NUREBASE: nuclear hormone receptors
  - NURSA: nuclear receptors
- OGRe - Organellar Genome Retrieval: mitochondrial genomes
- PSORTdb: protein subcellular localizations: ePSORT, cPSORT
- Secreted Protein Database: human, mouse and rat.
- THGS - Transmembrane Helices in Genome Sequences
Protein Databases

- General Sequence databases
- Protein properties
- Protein localization and targeting
- **Protein sequence motifs and active sites**
- Protein domain databases; protein classification
- Databases of individual protein families
Protein sequence motifs and active sites

- ASC - Active Sequence Collection
- Blocks
- CoC
- COMe - Co-Ordination of Metals etc.
- CoPS
- CSA - Catalytic Site Atlas
- eBLOCKS
- eF-site - Electrostatic surface of Functional site
- eMOTIF
- InterPro
- Metalloprotein Site Database
- O-GLYCBASE
- PDBSite
- PhosphoELM Base
- PRINTS
- PROMISE
- ProRule
- PROSITE
- ProTeus
- SitesBase
Protein Databases

• General Sequence databases
• Protein properties
• Protein localization and targeting
• Protein sequence motifs and active sites
• Protein domain databases; protein classification
• Databases of individual protein families
Protein domain databases; protein classification

- ADDA - Automatic Domain Decomposition Algorithm
- BAiBASE
- BIOZON
- CDD
- CluStr - Clusters of Swiss-Prot and TrEMBL proteins
- COG - Clusters of Orthologous Groups of proteins
- FunShift
- FusionDB
- Hits
- HSSP
- InterDom
- InterPro - PROSITE, Pfam, PRINTS, Prodom, SMART, TIGRFAMs, PIR superfamily
- iProClass
- MulPSSM
- PALI
- PANDIT
- Pfam
- PIRSF
- ProDom - SP, TrEMBL
- ProtoMap
- ProtoNet
- SBASE
- SIMAP
- SMART - Simple Modular Architecture Research Tool
- SUPFAM
- TCDB
- TIGRFAMs - HMM, GO annotations, MSA
- ProtRepeatsDB
Protein Databases

- General Sequence databases
- Protein properties
- Protein localization and targeting
- Protein sequence motifs and active sites
- Protein domain databases; protein classification
- Databases of individual protein families
Databases of individual protein families

- AARSDB
- ABCdb
- ARAMEMNON
- BacTregulators (formerly AraC/XylS database)
- CSDBase - Cold Shock Domain database
- DCCP - Database of Copper-Chelating Proteins
- DExH/D Family Database
- DSD
- Endogenous GPCR List
- EROP-Moscow
- ESTHER
- FUNPEP
- GPCRDB
- gpDB - G-protein database
- Histone Database
- HIV RT and Protease Sequence Database
- Homeobox Page
- Homeodomain Resource
- InBase
- King - Kinases in Genomes
- Knottins
- LGICdb
- Lipase Engineering Database
- Lipid MAPS
- LOX-DB
- MEROPS
- Nuclear Receptor Resource
- NucleaRDB
- NUREBASE
- NURSA
- Olfactory Receptor Database
- Peptaibol
- PHOTOPROT
- PLANT-PIs
- PlantsP/PlantsT
- PLPMDB
- ProLysED - Prokaryotic Lysis Enzymes Database
- Prolysis
- Protein kinase resource
- REBASE
- Ribonuclease P Database
- RINRdb
- RPG - Ribosomal Protein Gene database
- RTKdb - Receptor Tyrosine Kinase database
- SDAP
- SENTRA
- SEVENS
- SRPDB
- TransportDB
- VKCDB - Voltage-gated K+ Channel Database
- Wnt Database
Exercises

• Read NAR DB index site: search for different databases based on different search terms.
  – [http://www3.oup.co.uk/nar/database/c/](http://www3.oup.co.uk/nar/database/c/)

• Read uniprot manual at:

• Look for a sequence of your choice in GenPept and SwissProt. Study them w.r.t. no. of sequences your search yields, level of annotation, information etc. Do you notice any difference?
Exercises

• Read NAR DB paper and NAR DB index site: search for databases of your interest.

• Self study:
  – [http://www3.oup.co.uk/nar/database/c/](http://www3.oup.co.uk/nar/database/c/)
  – Study sequence retrieval tools at
    • [http://srs.ebi.ac.uk](http://srs.ebi.ac.uk)
    • [http://www.ebi.uniprot.org/index.shtml](http://www.ebi.uniprot.org/index.shtml)

• Study few derived databases for proteins
Database searching tips

• Look for links to Help or Examples
• Always check update dates
• Level of curation
• Try Boolean searches
• Be careful with UK/US spelling differences
  – leukaemia vs leukemia
  – haemoglobin vs hemoglobin
  – colour vs color