Bioinformatics Resources - Swissprot -

Lecture & Exercises

Prof. B. Rost, Dr. L. Richter, J. Reeb

Institut für Informatik I12
# Preliminary Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 13th</td>
<td>Intro, General Overview (0. sh.)</td>
<td>June 1st</td>
<td>Lecture cancelled</td>
</tr>
<tr>
<td>April 20th</td>
<td>Sequence Databases (1. sh.)</td>
<td>June 8th</td>
<td>NoSql 2 (6.sh.)</td>
</tr>
<tr>
<td>April 27th</td>
<td>Sequence Databases (2. sh.)</td>
<td>June 15th</td>
<td>MongoDB, JavaScript (7.sh.)</td>
</tr>
<tr>
<td>May 4th</td>
<td>Structure Databases (3. sh.)</td>
<td>June 22nd</td>
<td>Node.js Applications (8.sh.)</td>
</tr>
<tr>
<td>May 11th</td>
<td>Lecture cancelled</td>
<td>June 29th</td>
<td>PredictProtein</td>
</tr>
<tr>
<td>May 18th</td>
<td>SQL (4. sh.)</td>
<td>Jul 6th</td>
<td>Wrap Up, Q&amp;A</td>
</tr>
<tr>
<td>May 25th</td>
<td>SQL, NoSql (5. sh)</td>
<td>Jul 20th</td>
<td>Exam</td>
</tr>
</tbody>
</table>
Names and Other Complications

Amos Bairoch


Ioannis Xenarios

taken from http://www.isb-sib.ch/people/loannis.Xenarios

BioinfRes SoSe 18
History

1986 A. Bairoch created Swiss-Prot at the University of Geneva, since 1988 in collaboration with EMBL/EBI

1993 together with Ron Appel launch of ExPASy

1998 Foundation of SIB (Swiss Institute of Bioinformatics)

2002 Foundation of the UniProt consortium by EBI, SIB and PIR
UniProt Components:

- UniProtKB:
  - UniProtKB/Swiss-Prot
  - UniProtKB/TrEMBL
- UniParc: pure sequence archive, no annotations
- UniRef: consists of three databases of clustered sets of protein sequences (UniRef100, UniRef90, UniRef50) using the CD-HIT algorithm
- UniMes: data from metagenomic and environmental samples, not in UniProtKB
ExPASy

- Expert Protein Analysis System (1993)
- now: SIB ExPASy Bioinformatics Resources Portal
Expasy Categories

- Proteomics
- Genomics
- Structural Bioinformatics
- Systems biology
- Phylogeny/evolution
Expasy Categories

- Population genetics
- Transcriptomics
- Biophysics
- Imaging
- Drug Design
Resource Description

1. Resource name and description
2. Maintaining SIB group
3. Scientific category
4. Keywords: a controlled vocabulary is used to tag the resource
5. URL for the web interface and for the download if available

6. Software type: website, command line interface, GUI, etc

7. Status: green checkbox if currently available
UniProt/SwissProt Statistics

- Release 2018_04, Apr. 18\textsuperscript{th}
- 557,275 sequence entries, comprising 199,856,860 amino acids abstracted from 259,145 references
- in 2017_05: 554,515 sequence entries / 198509421 amino acids
What happens in an Update?

- Between 2018_3 und 2018_4:
  - 265 sequences have been added
  - 27 existing entries have been updated
  - annotation of 75954 entries have been revised
- Growth over three years: 2018_4 vs 2017_5 vs 2016_5 vs 2015_5

<table>
<thead>
<tr>
<th>Protein existence (PE)</th>
<th>Entries</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Evidence at protein level</td>
<td>98,566 (95,143) (92,536) (85,419)</td>
<td>17.7 (17.2) (16.8) (15.6)</td>
</tr>
<tr>
<td>2. Evidence at transcript level</td>
<td>57,060 (57,649) (57,757) (61,814)</td>
<td>10.2 (10.4) (10.5) (11.3)</td>
</tr>
<tr>
<td>3. Inferred from homology</td>
<td>386,164 (386,111) (387,589) (387,733)</td>
<td>69.3 (69.6) (70.3) (70.7)</td>
</tr>
<tr>
<td>4. Predicted</td>
<td>13,621 (13,751) (11,358) (11,526)</td>
<td>2.4 (2.5) (2.1) (2.1)</td>
</tr>
<tr>
<td>5. Uncertain</td>
<td>1864 (1,861) (1,953) (1,962)</td>
<td>0.3 (0.3) (0.4) (0.4)</td>
</tr>
</tbody>
</table>
Development

Number of entries in UniProtKB/Swiss-Prot


BioinfRes SoSe 18
More Numbers (rel. 2017_5)

- Represented species: 13425 (13.209 in 2015_5)
- Top 20 species: 119,149 sequences, i.e. 21.5% of the total number of entries

<table>
<thead>
<tr>
<th>Entries</th>
<th>No of Species</th>
<th>Entries</th>
<th>No of Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5,570 (5,495)</td>
<td>8</td>
<td>240 (228)</td>
</tr>
<tr>
<td>2</td>
<td>1,921 (1,899)</td>
<td>9</td>
<td>229 (214)</td>
</tr>
<tr>
<td>3</td>
<td>1,054 (1,023)</td>
<td>10</td>
<td>133 (122)</td>
</tr>
<tr>
<td>4</td>
<td>668 (657)</td>
<td>11-20</td>
<td>719 (711)</td>
</tr>
<tr>
<td>5</td>
<td>493 (487)</td>
<td>21-50</td>
<td>442 (426)</td>
</tr>
<tr>
<td>6</td>
<td>405 (399)</td>
<td>51-100</td>
<td>217 (213)</td>
</tr>
<tr>
<td>7</td>
<td>285 (289)</td>
<td>&gt;100</td>
<td>1049 (1.046)</td>
</tr>
</tbody>
</table>
### Species Representation (rel. 2017_5/ 2015_5)

<table>
<thead>
<tr>
<th>Top</th>
<th>Frequency</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20,201 (+3)</td>
<td>Homo sapiens (Human)</td>
</tr>
<tr>
<td>2</td>
<td>16,877 (+166)</td>
<td>Mus musculus (Mouse)</td>
</tr>
<tr>
<td>3</td>
<td>15,333 (+1,445)</td>
<td>Arabidopsis thaliana (Mouse-ear cress)</td>
</tr>
<tr>
<td>4</td>
<td>7,989 (+68)</td>
<td>Rattus norvegicus (Rat)</td>
</tr>
<tr>
<td>5</td>
<td>6,721 (+3)</td>
<td>Saccharomyces cerevisiae (Baker’s yeast)</td>
</tr>
<tr>
<td>6</td>
<td>5,999 (+6)</td>
<td>Bos taurus (Bovine)</td>
</tr>
<tr>
<td>7</td>
<td>5,141 (+38)</td>
<td>Schizosaccheromyces pombe (Fission yeast)</td>
</tr>
<tr>
<td>8</td>
<td>4,435 (+2)</td>
<td>Escherichia coli K12</td>
</tr>
<tr>
<td>9</td>
<td>4,185 (+0)</td>
<td>Bacillus subtilis</td>
</tr>
<tr>
<td>10</td>
<td>4,134 (+3)</td>
<td>Dictyostelium discoideum (Slime mold)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Representation of the Divisions (rel. 2017_5)

- Bacteria (60%), 333,047
- Eukaryota (33%), 185,305
- Viruses (3%), 16,719
- Archaea (4%), 19444
Distribution of Eukaryota (rel. 2017_5)

- Human (11%), 20,202
- Other Mammalia (25%), 46,449
- Other (8%), 15,339
- Viridiplantae (21%), 38,739
- Other Vertebrata (10%), 18,066
- Fungi (18%), 32,674
- Insecta (5%), 9,059
- Nematoda (3%), 4,777
Length Distribution (rel. 2017_5)
Amino Acid Composition (rel. 2017_5)

gray=aliphatic, red=acidic, green=small hydroxy, blue=basic, black=aromatic, white=amide, yellow=sulfur

BioinfRes SoSe 18
SwissProt Annotation Process

- defined in
  http://www.uniprot.org/docs/sop_manual_curation.pdf

- explained in
  http://www.uniprot.org/help/manual_curation
Annotation Phases

1. Sequence curation
2. Sequence analysis
3. Literature curation
4. Family-based curation
5. Evidence attribution
6. Quality assurance, integration and update
Sequence Curation

- more than 95% are translated CDS from INSDC
- other sources: PDB, direct protein sequencing, projects not submitting to INSDC
- sequences are selected according to curation priorities (http://www.uniprot.org/program/)
- results in the “canonical sequence” for a gene/species pair
Steps toward the canonical sequence

- Entry selection
- Run BLAST similarity searches to identify additional sequences for the same gene
- Identify homologs by reciprocal BLAST and phylogeny based resources
- Lock selected entries for other curators to prevent duplication
Steps toward the canonical sequence

- Prepare sequence alignments with T-Coffee, Muscle, Clustal W
- Merge into the canonical sequence:
  - most prevalent
  - most similar to orthologs sequences found in other species
  - based on length and aa composition it allows the clearest description
  - default: longest
- record conflicts and variations
Sequence Analysis

- Several analysis programs are applied to the sequences for:
  - topological features
  - post-translational modifications
  - domains

- All results are manually checked and included or excluded for annotation.
## Topological Analysis

<table>
<thead>
<tr>
<th>Tools</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal P</td>
<td>Presence and location of signal peptides</td>
</tr>
<tr>
<td>TargetP</td>
<td>Presence and location of transit peptides</td>
</tr>
<tr>
<td>Predotar</td>
<td>Mitochondrial, plastid or ER targeting sequences</td>
</tr>
<tr>
<td>ESKW</td>
<td>Transmembrane domains</td>
</tr>
<tr>
<td>MEMSAT</td>
<td>Transmembrane domains</td>
</tr>
<tr>
<td>TMHMM</td>
<td>Transmembrane domains</td>
</tr>
<tr>
<td>Phobius</td>
<td>Discriminates transmembrane and signal regions</td>
</tr>
</tbody>
</table>
Post-translational modification Analysis

<table>
<thead>
<tr>
<th>Tools</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPI-predictor</td>
<td>GPI lipid anchor sites</td>
</tr>
<tr>
<td>NetNGlyc</td>
<td>N-glycosylation sites</td>
</tr>
<tr>
<td>NetOGlyc</td>
<td>O-glycosylation sites</td>
</tr>
<tr>
<td>NMT Predictor</td>
<td>N-terminal myristoylation sites</td>
</tr>
<tr>
<td>Sulfinator</td>
<td>Tyrosine sulfatation sites</td>
</tr>
</tbody>
</table>
## Domain Analysis

<table>
<thead>
<tr>
<th>Tools</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>ps_scan</td>
<td>internal PROSITE profile, pattern and rule scanning</td>
</tr>
<tr>
<td>InterPro</td>
<td>retrieves non-PROSITE motif matches using InterPro database or InterProScan</td>
</tr>
<tr>
<td>Coils</td>
<td>Coiled-coils regions</td>
</tr>
<tr>
<td>polyAA</td>
<td>internal program which identifies homopolymeric stretches of amino acids</td>
</tr>
<tr>
<td>REPEAT</td>
<td>identifies the following repeats: Ankyrin, Armadillo, HAT, HEAT, Kelch, Leucine-rich, PFTA, PFTB, RCC1, TPR, WD40</td>
</tr>
</tbody>
</table>
Automatically selected results are returned in a graphical interface which allows visualisation of the predictions (Figure 1). Selected features are shown in green and unselected features are shown in red. The selected/unselected state of a feature can be toggled by clicking on it.

Figure 1. UniProtKB sequence analysis results displayed in graphical interface.

All predictions are manually reviewed and relevant results are selected for inclusion in the entry. The sequence analysis platform then transforms the selected features into UniProtKB annotation by applying a set of automatic annotation rules (Figure 2).

Literature Curation

- Identification of relevant scientific literature from:
  - literature and text mining resources (PubMed, Europe PMC, iHOP, TextPresso)
  - additions from other sources made by the curator

- Information is extracted from the full text:
  - general annotations (not position specific)
  - position specific annotations
General Annotations

- http://www.uniprot.org/help/general_annotation
- position-independent
- contains mostly general biological information like: functions, catalytic activity, cofactor, enzyme regulation, subunit structure, pathway,...
Sequence Annotations

- position dependent
- [http://www.uniprot.org/help/sequence_annotation](http://www.uniprot.org/help/sequence_annotation)
- regions or sites of interest like post-translational modifications, binding sites, active sites, etc.
- contains several subsections: molecule processing, regions, sites, amino acid modifications, natural variants, experimental info, secondary structure
Family-based Curation

- Evaluation and curation of homologs as described above
- Standardization of annotation of homologs
- Propagation of annotation across the homologs to ensure consistency
Evidence Attribution

- Every annotation is attributed to its original source
- Every annotation can be traced back and evaluated
- For evidence distinction there are 7 codes from the Evidence Code Ontology (ECO) used for manually curated entries
  
  - http://www.uniprot.org/help/evidences
  - Additional GO term annotation
<table>
<thead>
<tr>
<th>ECO code</th>
<th>Term name</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECO:0000269</td>
<td>experimental evidence used in manual assertion</td>
<td>Information for which there is published experimental evidence</td>
</tr>
<tr>
<td>ECO:0000303</td>
<td>non-traceable author statement used in manual assertion</td>
<td>Information based on author statements in scientific articles for which there is no experimental support</td>
</tr>
<tr>
<td>ECO:0000250</td>
<td>sequence similarity evidence used in manual assertion</td>
<td>Information which has been propagated from a related experimentally characterised protein</td>
</tr>
<tr>
<td>ECO:0000312</td>
<td>imported information used in manual assertion</td>
<td>Information which has been imported from another database and manually verified</td>
</tr>
<tr>
<td>ECO:0000305</td>
<td>curator inference used in manual assertion</td>
<td>Information which has been inferred by a curator based on his/her scientific knowledge or on the scientific content of an article</td>
</tr>
<tr>
<td>ECO:0000255</td>
<td>match to sequence model evidence used in manual assertion</td>
<td>Information originating from the UniProt automatic annotation systems or any of the sequence analysis programs used during the manual curation process and which has been manually verified</td>
</tr>
<tr>
<td>ECO:0000244</td>
<td>combinatorial evidence used in manual assertion</td>
<td>Information which is manually curated based on a combination of experimental and computational evidence</td>
</tr>
</tbody>
</table>

Quality Control and Integration

- Finished entries run through a series of rule-based checked concerning especially positions and regions
- All errors are corrected
- Manually reviewed by a senior curator
- Finally it is integrated into the database
- Unlock the finished entries for further curation
Demostration

- [http://www.uniprot.org/uniprot/P62756#section_features](http://www.uniprot.org/uniprot/P62756#section_features)
The Swiss-Prot Flat File

- An entry is composed by different line types
- Line types have their own format
- Follows EMBL Nucleotide Sequence Database format as close as possible
- 2 sections:
  - core data (sequence data, citation info, taxonomy)
  - annotations (function, modification, domains, secondary and quart structure, disease associations, conflicts, etc.)
The following table lists the available two-letter line codes. Each code is followed by three blanks.

<table>
<thead>
<tr>
<th>Line Code</th>
<th>Content</th>
<th>Occurrence in an entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Identification</td>
<td>Once; starts the entry</td>
</tr>
<tr>
<td>AC</td>
<td>Accession number(s)</td>
<td>Once or more</td>
</tr>
<tr>
<td>DT</td>
<td>Date</td>
<td>Three times</td>
</tr>
<tr>
<td>DE</td>
<td>Description</td>
<td>Once or more</td>
</tr>
<tr>
<td>GN</td>
<td>Gene name(s)</td>
<td>Optional</td>
</tr>
<tr>
<td>OS</td>
<td>Organism species</td>
<td>Once or more</td>
</tr>
<tr>
<td>OG</td>
<td>Organelle</td>
<td>Optional</td>
</tr>
<tr>
<td>OC</td>
<td>Organism classification</td>
<td>Once or more</td>
</tr>
<tr>
<td>OX</td>
<td>Taxonomy cross-reference</td>
<td>Once</td>
</tr>
<tr>
<td>OH</td>
<td>Organism host</td>
<td>Optional</td>
</tr>
</tbody>
</table>

--continued--
<table>
<thead>
<tr>
<th>Line Code</th>
<th>Content</th>
<th>Occurence in an entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>RN</td>
<td>Reference number</td>
<td>Once or more</td>
</tr>
<tr>
<td>RP</td>
<td>Reference position</td>
<td>Once or more</td>
</tr>
<tr>
<td>RC</td>
<td>Reference comment(s)</td>
<td>Optional</td>
</tr>
<tr>
<td>RX</td>
<td>Reference cross-reference(s)</td>
<td>Optional</td>
</tr>
<tr>
<td>RG</td>
<td>Reference group</td>
<td>Once or more (Optional if RA line)</td>
</tr>
<tr>
<td>RA</td>
<td>Reference authors</td>
<td>Once or more (Optional if R line)</td>
</tr>
<tr>
<td>RT</td>
<td>Reference title</td>
<td>Optional</td>
</tr>
<tr>
<td>RL</td>
<td>Reference location</td>
<td>Once or more</td>
</tr>
<tr>
<td>CC</td>
<td>Comments or notes</td>
<td>Optional</td>
</tr>
<tr>
<td>DR</td>
<td>Database cross-references</td>
<td>Optional</td>
</tr>
<tr>
<td>PE</td>
<td>Protein existence</td>
<td>Once</td>
</tr>
<tr>
<td>KW</td>
<td>Keywords</td>
<td>Optional</td>
</tr>
<tr>
<td>FT</td>
<td>Feature table data</td>
<td>Once or more in Swiss-Prot, optional in TrEMBL</td>
</tr>
<tr>
<td>SQ</td>
<td>Sequence header</td>
<td>Once</td>
</tr>
<tr>
<td>(blanks)</td>
<td>Sequence data</td>
<td>Once or more</td>
</tr>
<tr>
<td>//</td>
<td>Termination line</td>
<td>Once; ends the entry</td>
</tr>
</tbody>
</table>
Fields in More Detail

- **ID line:**
  - ID  *EntryName Status; SequenceLength*.

- **EntryName:** up to 11 uppercase alphanumeric characters \(X_Y\)
  - \(X\) is a mnemonic code of at most 5 alphanumeric characters
  - \(Y\) is a mnemonic species identification code of at most 5 alphanumeric characters

- **ID**  CYC_BOVIN  Reviewed;  104 AA.

BioinfRes SoSe 18
• AC line:
  AC  AC_number_1;[ AC_number_2;]...[ AC_number_N;]

• Accession number: 6 or 10 characters

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>[A-N,R-Z]</td>
<td>[0-9][A-Z]</td>
<td>[A-Z, 0-9]</td>
<td>[A-Z, 0-9]</td>
<td>[0-9]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[O,P,Q]</td>
<td>[0-9][A-Z, 0-9]</td>
<td>[A-Z, 0-9]</td>
<td>[A-Z, 0-9]</td>
<td>[0-9]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[A-N,R-Z]</td>
<td>[0-9][A-Z]</td>
<td>[A-Z, 0-9]</td>
<td>[A-Z, 0-9]</td>
<td>[0-9][A-Z]</td>
<td>[A-Z,0-9]</td>
<td>[A-Z,0-9]</td>
<td>[0-9]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• RegEx: [OPQ][0-9][A-Z0-9]{3}[0-9] | [A-NR-Z][0-9]
  ([A-Z][A-Z0-9]{2}[0-9]){1,2}

• Examples: P12345, Q1AAA9, A0A022YWF9
● DT line: date, DD-MMM-YYYY
● always one of the biweekly release dates
● always three lines:
  - date of integration
  - date of sequence version, sequence version X
  - date of entry version, entry version X
● Example:

DT  01-FEB-1999, integrated into UniProtKB/TrEMBL.
DE lines:

- three categories and additional subcategories
- contains a recommended name
- besides: full name, short name, EC number
- alternative names: e.g. as an allergen or in biotechnology, ...
DE RecName: Full=Annexin A5;
DE Short=Annexin-5;
DE AltName: Full=Annexin V;
DE AltName: Full=Lipocortin V;
DE AltName: Full=Endonexin II;
DE AltName: Full=Calphobindin I;
DE AltName: Full=CBP-I;
DE AltName: Full=Placental anticoagulant protein I;
DE Short=PAP-I;
DE AltName: Full=PP4;
DE AltName: Full=Thromboplastin inhibitor;
DE AltName: Full=Vascular anticoagulant-alpha;
DE Short=VAC-alpha;
DE AltName: Full=Anchorin CII;
DE RecName: Full=Granulocyte colony-stimulating factor;
DE Short=G-CSF;
DE AltName: Full=Pluripoiyetin;
DE AltName: Full=Filgrastim;
DE AltName: Full=Lenograstim;
DE Flags: Precursor;
OS line: originating organism

OS  Homo sapiens (Human).

OS  Rous sarcoma virus (strain Schmidt-Ruppin A) (RSV-SRA) (Avian leukosis OS  virus-RSA).

OC lines: contain the taxonomic classification of the source organism according to (http://www.ncbi.nlm.nih.gov/Taxonomy/)

OC  Node[; Node...].

OC  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC  Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae;
OC  Homo.
RN, RP, RC, RX, RG, RA, RT, RL

- can occur multiple time
- order in block fixed

E.g:
- RN [1]
- RP NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS A AND C), FUNCTION, INTERACTION
- RP WITH PKC-3, SUBCELLULAR LOCATION, TISSUE SPECIFICITY, DEVELOPMENTAL
- RP STAGE, AND MUTAGENESIS OF PHE-175 AND PHE-221.
- RC STRAIN=Bristol N2;
- RX PubMed=11134024; DOI=10.1074/jbc.M008990200;
- RT "A novel adapter protein employs a phosphotyrosine binding domain and
  exceptiona) basic N-terminal domains to capture and localize an
  atypical protein kinase C: characterization of Caenorhabditis elegans
  C kinase adapter 1, a protein that avidly binds protein kinase C3."
CC lines

- free text
- contains most of the annotated information
  - CC  -!- TOPIC: First line of a comment block;
    CC second and subsequent lines of a comment block.
- structured by predefined topics like: Allergen, Alternate Products, ..., Cofactor, ..., Disease, .. Domain,..., Function, Interaction, .......
ALLERGEN: Causes an allergic reaction in human. Minor allergen of bovine dander.

ALTERNATIVE PRODUCTS:
Event=Alternative initiation; Named isoforms=2;
Name=Alpha;
IsoId=P51636-1; Sequence=Displayed;
Name=Beta;
IsoId=P51636-2; Sequence=VSP_018696;

SUBCELLULAR LOCATION: Cell membrane {ECO:0000250}; Peripheral membrane protein {ECO:0000250}. Secreted {ECO:0000250}. Note=The last 22 C-terminal amino acids may participate in cell membrane attachment.

SUBCELLULAR LOCATION: Isoform 2: Cytoplasm {ECO:0000305}. 
Cross References

- too many to enumerate
- extensive references with nucleotide databases,

\[\text{e.g.:}\]
\begin{verbatim}
in EMBL
FT  CDS 302..2674
FT  /protein_id="CAA03857.1"
FT  /db_xref="SWISS-PROT:P26345"
FT  /gene="recA"
FT  /product="RecA protein"
\end{verbatim}

\begin{verbatim}
in Swiss=Prot
DR  EMBL; AJ297977; CAC17465.1; -; Genomic_DNA.
DR  EMBL; X56491; CAA39846.1; ALT_FRAME; mRNA.
\end{verbatim}
Key Words / Feature Table

- KW  Keyword[; Keyword...].
- helps to search resp. index the database
- no limits:
  KW  3D-structure; Alternative splicing; Alzheimer disease; Amyloid;
  KW  Apoptosis; Cell adhesion; Coated pits; Copper;
  KW  Direct protein sequencing; Disease mutation; Endocytosis;
  KW  Glycoprotein; Heparin-binding; Iron; Membrane; Metal-binding;
  KW  Notch signaling pathway; Phosphorylation; Polymorphism;
  KW  Protease inhibitor; Proteoglycan; Serine protease inhibitor; Signal;
  KW  Transmembrane; Zinc.

- Feature table like GenBank/EMBL/DDBJ
Programmatic Access

- [http://www.uniprot.org/help/programmatic_access](http://www.uniprot.org/help/programmatic_access) (remember this link!)
- several use cases documented, but not as an API
- best way: use the web interface to construct/refine your query first before you try to automate the process
Retrieving an Individual Entry

- uses simple URL which can be bookmarked
- for individual entries: http://www.uniprot.org/uniprot/P12345
- default result is a web page
- alternative formats: txt, xml, rdf, fasta, gff
- specified via the accession suffix
- structured formats like xml or rdf can include referenced entries
Using the ID mapping service

- [http://www.uniprot.org/help/programmatic_access#batch_retrieval_perl_example](http://www.uniprot.org/help/programmatic_access#batch_retrieval_perl_example)
- uses http POST method
- converts between different database IDs
- you have to know the specific abbreviation for the respective databases
Retrieving Entries via Queries

- uses http GET method i.e.
- the query string is part of the URL
- structure might be quite complex
- use the browser to configure the query string
- more setting are available via the query builder
  http://www.uniprot.org/help/advanced_search
- the URL length might be limited to 1000 characters
Examples

- http://www.uniprot.org/uniprot/P12345.txt
- http://www.uniprot.org/uniprot/P12345.xml
- http://www.uniprot.org/uniref/UniRef90_P04259.xml
- http://www.uniprot.org/uniref/UniRef90_P04259.rdf
- http://www.uniprot.org/uniref/UniRef90_P04259.fasta
- http://www.uniprot.org/uniref/UniRef90_P04259.tab