The UniProtKB/Swiss-Prot Human Proteomic Initiative (HPI) as a support for the characterization of human proteome in health and disease
Swiss-Prot was created in July 1986. Since 1987, it is a collaboration between the SIB and the EBI; from 2003 onward it is the central part of the UniProt project (collaboration between SIB, EBI and PIR).

- It is an annotated, non-redundant, cross-referenced, documented protein sequence knowledge resource, which contains 255'000 sequences; 130'000 literature references; 3'500'000 cross-references to 100 databases; ~600 Mb of annotations.

- About 3'600'000 sequences are available in TrEMBL, Swiss-Prot’s computer-annotated supplement. TrEMBL contains all the submitted coding sequences (CDS) from EMBL, translated but not yet integrated in Swiss-Prot.

- The UniProt Knowledgebase which contains both UniProtKB/Swiss-Prot and UniProtKB/TrEMBL is released every other week and available from about 50 servers, the main source being ExPASy (www.expasy.org, www.uniprot.org).
CDS: CoDing Sequence (CDS) provided by the submitters

Transcription start

join 397..627, 1194..1339, 1596..1682, 2294..2473, 2608..3327)

Translation start

CDS translation provided by EMBL

integrated in TrEMBL
Automated extraction of protein sequence (CDS), gene name and references + Automated annotation
In UniProtKB/Swiss-Prot, 1 entry = 1 gene

i) Merge of all known protein sequences (CDS) derived from the same gene

-> avoid redundancy and improve sequence reliability

(for human: ~ 6 different sequence reports per entry)

ii) Annotation of the sequence differences
    (including conflicts, polymorphisms, splice variants etc..)

-> annotation of protein diversity
10 former TrEMBL entries were merged into a single Swiss-Prot entry.

13 CDS were used to validate the final sequence.

Among these CDS, there is one sequence conflict.
And soon the new ‘Protein existence evidence’ tag
The goals of the Swiss-Prot Human Proteome Initiative

• **Annotation of all known human proteins and their polymorphisms**
  ➔ We plan to complete the first round of the human proteome annotation in *September 2008*; this means annotating about 5’000 new entries while keeping up with the annotation of the 15’500 existing ones;

• **Annotation of mammalian orthologs of human proteins**
  ➔ We will attempt to keep up with the annotation of the mouse proteome and significantly speed up that of the rat;
  ➔ We also hope to be able to start the use of automated annotation procedure to propagate annotation to newly sequenced mammalian proteome sets.
Protein polymorphism annotation

- Swiss-Prot already holds information on almost **31'000** protein variants;

- Only non-synonymous ‘**c-SNPs**’ (coding single nucleotide polymorphisms), also called ‘**SAPs**’ (single amino-acid polymorphisms) are annotated

- About **50%** of them are linked to genetic disorders.

---

**Note:** Mutations that cause major changes to a protein sequence (such as frameshift mutations) are not considered to be relevant to Swiss-Prot, as their deleterious effect on a given protein’s function is usually obvious.
Disease-related annotation

• According to OMIM, there are currently 2'000 known genes associated with human genetic disorders;
• Swiss-Prot contains disease-related information for about 2'100 proteins. Among them, 1618 are linked to an OMIM description of a phenotype.
• 1'443 Swiss-Prot entries contain information on sequence variants associated with a disease state (i.e. have the keyword «disease mutation»). When possible, additional «medical-oriented» keywords are provided:

<table>
<thead>
<tr>
<th>Natural variations</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Splice variant 311 – 422</td>
<td>112</td>
</tr>
<tr>
<td>Splice variant 311 – 359</td>
<td>48</td>
</tr>
<tr>
<td>Sequence variant 65</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 248</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 249</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 250</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 322</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 370</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 371</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 372</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 375</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 380</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 384</td>
<td>1</td>
</tr>
<tr>
<td>Sequence variant 391</td>
<td>1</td>
</tr>
</tbody>
</table>

Defects in FGFR3 are the cause of achondroplasia (ACH) [MIM: 100900]. ACH is an autosomal dominant disease and is the most frequent form of short-limb dwarfism. It is characterized by a long, narrow trunk, short extremities, particularly in the proximal (mizomelic) segments, a large head with frontal bossing, hypertelorism of the midfacial and a short configuration of the hands.

Defects in FGFR3 are a cause of Crouzon syndrome [MIM: 123500], also called craniofacial dysostosis type I (CFD1). Crouzon syndrome is characterized by craniosynostosis (premature fusion of the skull sutures), hypertelorism, exophthalmos and external strabismus, pear-shaped nose, short upper lip, hypertelorism maxilla, and a relative mandibular prognathism.

Defects in FGFR3 are the cause of platyspondylic lethal skeletal dysplasia San Diego type (PLSD-SD) [MIM: 270230]. Platyspondylic lethal skeletal dysplasias (PLSDs) are a heterogeneous group of chondrodysplasias characterized by severe platyspondyly and limb shortening. PLSD-SD is characterized by postnatal growth deficiency, mild developmental delay, short trunk, craniofacial abnormalities, platyspondyly, delayed ossification, generalized osteoporosis and thin ribs.
How many human proteins?

**Genome**

~ 21'000 human genes

Alternative splicing of mRNA

2-5 fold increase

**Transcriptome**

~ 80 to 100’000 human transcripts

**Proteome**

~ 1’000’000 human proteins

Post-translational modifications of proteins (PTMs)

5-10 fold increase
Post-translational modifications (PTMs)

- Sequences are important, but are generally not fully representative of the final ‘biological entity’: **most proteins are the target of PTMs**.

- **PTMs are important at various levels**, including the 3D structure, interactions, subcellular location and also the function.

- **283 different standardized PTM descriptions are currently available in Swiss-Prot** (excluding processing, disulfide bonds and glycosylation events).
This document is/will be used by several protein identification software
Proteomics help to characterize the human proteome by:

- confirming the *in vivo* existence of potential splice isoforms and their subcellular/tissular location
- resolving sequence conflicts, confirming the existence of particular variants
- localizing translation initiation sites and enzymatic cleavage sites
- allowing the characterization of protein complexes
- identifying/confirming PTMs and addressing their fluctuations over time and space.
Protein identification and database accuracy are interdependent

Proteomics experiment

Database search

Sequence correction, Annotation enrichment (PTM, splicing forms)

Data filtering, curation

UniProt

Direct submission (identifications only)

Protein identifications

Data filtering, curation

Publication

Phosphoproteome analysis of the human mitotic spindle

Marisana Nouskillera, Herman H. W. Stijn, Golde Sauerb, Erich A. Migo, and Roman Körner

Submission to specialized proteomic databases (spectra+identifications)
Proteomic studies already allowed to update 2767 human Swiss-Prot entries, mainly with PTM information

(UniProtKB release 10.0, March 2007)

Phosphorylation (83%)  
Glycosylation (9%)  
Subcellular location (4%)  
Other PTMs (4%)
HPI Status report

UniProtKB/Swiss-Prot Release 51.7 of 20-Feb-2007: 259034 entries (Release statistics)

Total number of manually annotated mammalian sequences in UniProtKB/Swiss-Prot: 50,075

Total number of manually annotated HUMAN sequences in UniProtKB/Swiss-Prot: 15,727

<table>
<thead>
<tr>
<th>Annotation categories</th>
<th>counts [total / unique]</th>
<th>max per entry</th>
<th>average per entry</th>
<th>number of entries</th>
<th>coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additional isoforms due to alternative splicing, initiation or promoter usage:</td>
<td>9,189 / 9,189</td>
<td>32</td>
<td>0.58</td>
<td>4,888</td>
<td>31%</td>
</tr>
<tr>
<td>Variants (disease mutations and polymorphisms):</td>
<td>30,531 / 30,465</td>
<td>464</td>
<td>1.94</td>
<td>5,454</td>
<td>34%</td>
</tr>
<tr>
<td>Post-translational modifications (experimentally proven or potential):</td>
<td>46,109 / 22,645</td>
<td>303</td>
<td>2.93</td>
<td>8,528</td>
<td>54%</td>
</tr>
<tr>
<td>References to published articles:</td>
<td>76,693 / 38,165</td>
<td>156</td>
<td>4.88</td>
<td>15,526</td>
<td>98%</td>
</tr>
<tr>
<td>Comment blocks:</td>
<td>86,043 / 33,522</td>
<td>30</td>
<td>5.47</td>
<td>15,264</td>
<td>97%</td>
</tr>
<tr>
<td>Feature lines:</td>
<td>375,257 / 275,393</td>
<td>1046</td>
<td>23.86</td>
<td>15,727</td>
<td>100%</td>
</tr>
</tbody>
</table>

Cross-references:

- EMBL protein_ids: 100,398 / 100,227
- InterPro
- PDB (3D-structure)
- MIM catalog of human genes and genetic disorders: 12,680 / 11,980
- HGNC gene nomenclature: 15,189 / 15,040

+ 1480 cross-references to Human Protein Atlas in 1320 entries
From pull to push..

- For now more than 20 years we have been «pulling» information and knowledge from various sources, but mainly from literature;

- It is now time to make sure that the next 20 years will be defined by the fact that researchers «push» their results and the interpretation of their results in the knowledgebase.
If you have this type of results (preferably already submitted to a proteomics data repository)

Importance of submitting your MS/MS data to databases...

All peptides listed below were manually controlled on the spectra. Even though some scores are low, they are all assumed to be correct and don’t have to be questioned.

The following series of rat peptides were obtained by MS/MS after tryptic digestion:

<table>
<thead>
<tr>
<th>Start - End</th>
<th>Observed</th>
<th>Mr(calc)</th>
<th>Delta</th>
<th>Miss-Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 - 8</td>
<td>429.7833</td>
<td>857.5520</td>
<td>0.0090</td>
<td>0 M.AQILPIR.F N-Acetyl (Protein): 13C6 (R) (Ions score 44)</td>
</tr>
<tr>
<td>87 - 96</td>
<td>526.6577</td>
<td>1251.0945</td>
<td>1231.6533</td>
<td>0 K.TEQUFLMK.S Oxidation (M) (Ions score 29)</td>
</tr>
<tr>
<td>177 - 188</td>
<td>669.3645</td>
<td>1336.7145</td>
<td>1336.6809</td>
<td>0 R.VQAMQLYSVD.K (Ions score 47)</td>
</tr>
<tr>
<td>199 - 205</td>
<td>615.6514</td>
<td>1283.3025</td>
<td>1283.9500</td>
<td>0 K.KVQPIEGKAASFAQK.K (Ions score 37)</td>
</tr>
<tr>
<td>200 - 205</td>
<td>572.6672</td>
<td>1145.3797</td>
<td>1145.8631</td>
<td>0 K.KVQPIEGKAASFAQK.K (Ions score 52)</td>
</tr>
<tr>
<td>270 - 278</td>
<td>536.3214</td>
<td>1070.6282</td>
<td>1070.6124</td>
<td>0 K.HDVFLITK.Y (Ions score 31)</td>
</tr>
<tr>
<td>355 - 366</td>
<td>652.8556</td>
<td>1303.6967</td>
<td>1303.6520</td>
<td>0 R.NNLAGAEELFAR.K (Ions score 58)</td>
</tr>
<tr>
<td>367 - 382</td>
<td>879.9643</td>
<td>1757.9140</td>
<td>1757.8736</td>
<td>0 K.KFNALFQGNYSEAAK.V (Ions score 57)</td>
</tr>
<tr>
<td>838 - 851</td>
<td>776.4073</td>
<td>1550.8001</td>
<td>1550.7464</td>
<td>0 R.GQFSITDELWAEVEK.R (Ions score 47)</td>
</tr>
<tr>
<td>892 - 892</td>
<td>667.0446</td>
<td>1333.6751</td>
<td>1333.6252</td>
<td>0 K.IYIDSNMNNPER.F (Ions score 52)</td>
</tr>
<tr>
<td>896 - 903</td>
<td>522.2309</td>
<td>1042.4623</td>
<td>1042.4356</td>
<td>0 R.ENFYDGR.V (Ions score 30)</td>
</tr>
<tr>
<td>1011 - 1022</td>
<td>708.3926</td>
<td>1414.7711</td>
<td>1414.7204</td>
<td>0 K.IVLKNSVYSEHR.N (Ions score 79)</td>
</tr>
<tr>
<td>1023 - 1034</td>
<td>677.4335</td>
<td>1352.8525</td>
<td>1352.8391</td>
<td>0 R.NCONULLLTAIAK (Ions score 34)</td>
</tr>
<tr>
<td>1095 - 1101</td>
<td>443.2231</td>
<td>884.4356</td>
<td>884.4028</td>
<td>0 K.ATEFAR.C (Ions score 26)</td>
</tr>
<tr>
<td>1123 - 1130</td>
<td>469.7484</td>
<td>937.4823</td>
<td>937.4756</td>
<td>0 K.EAZDIAK (Ions score 44)</td>
</tr>
<tr>
<td>1216 - 1226</td>
<td>651.8677</td>
<td>1301.7208</td>
<td>1301.6823</td>
<td>0 K.LVTHNSNGFR.L 13C6 (R) (Ions score 52)</td>
</tr>
<tr>
<td>1398 - 1406</td>
<td>599.8094</td>
<td>1097.6042</td>
<td>1097.5757</td>
<td>0 K.VANPELYYK.A (Ions score 51)</td>
</tr>
<tr>
<td>1435 - 1441</td>
<td>414.7243</td>
<td>827.4341</td>
<td>827.4177</td>
<td>0 K.RNYPVS.K (Ions score 54)</td>
</tr>
<tr>
<td>1444 - 1453</td>
<td>616.9137</td>
<td>1231.8129</td>
<td>1231.7748</td>
<td>0 K.OPLUVFYLR.S 13C6 (R) (Ions score 29)</td>
</tr>
<tr>
<td>1502 - 1508</td>
<td>472.2770</td>
<td>942.5395</td>
<td>942.4923</td>
<td>0 K.MELIEFR.R (Ions score 22)</td>
</tr>
<tr>
<td>1536 - 1545</td>
<td>573.2504</td>
<td>1144.4862</td>
<td>1144.4706</td>
<td>0 K.DAMQYASESF.D Oxidation (M) (Ions score 55)</td>
</tr>
<tr>
<td>1610 - 1630</td>
<td>610.9584</td>
<td>1227.9063</td>
<td>1227.6600</td>
<td>0 K.VKRLDASEELRK 13C6 (R) (Ions score 43)</td>
</tr>
</tbody>
</table>
... Or if you have new information about protein localisation of function...

... please, let us know!

swiss-prot@expasy.org

UniProtKB/Swiss-Prot Entry Q96GD4 (AURKB_HUMAN)★
Last modified February 20, 2007. Version 71. History...

Corresponding updates in the UniProtKB/Swiss-Prot entry will be done in priority!
UniProt curators participating to the HPI initiative

Amos Bairoch (group leader)
Silvia Jimenez (coordinator)
Yasmin Alam
Ghislaine Argoud Puy
Cecilia Arighi
Marie-Claude Blatter
Silvia Braconi
Lionel Breuza
Alan Bridge
Wei Mun Chan
Danielle Coral
Isabelle Cusin
Anne Estreicher
Livia Famiglietti
Serenella Ferro
Gabriella Frigerio
Arnaud Gos
Nadine Gruaz-Gumowski
Ursula Hinz
Lydie Lane
Michele Magrane
Madelaine Moinat
Sandra Ochard
Sylvain Poux
Sorogini Reynaud
Bernd Röchert
André Stutz
Shyamala Sundaram

A typical day in the life of a modern bioinformatician

Thank you!