Why study proteins

• Proteins are the work horses of the cell – and it is proteins which are the targets of therapeutic agents

• Knowing what proteins are being expressed in your healthy cells and how this alters in disease is critical to understanding the disease process

• Proteins cannot be studied in isolation – they need to be seen in the context of the network of interactions they make in the cell
Importance of reference protein sequence databases

• **Completeness and minimal redundancy**
A non redundant protein sequence database, with maximal coverage including splice isoforms, disease variant and PTMs.
Low degree of redundancy for facilitating peptide assignments

• **Stability and consistency**
Stable identifiers and consistent nomenclature
Databases are in constant change due to a substantial amount of work to improve their completeness and the quality of sequence annotation

• **High quality protein annotation**
Detailed information on protein function, biological processes, molecular interactions and pathways cross-referenced to external source
UniProtKB

- UniProt Knowledgebase:
  - 2 sections
  1. UniProtKB/Swiss-Prot Non-redundant, high-quality manual annotation – reviewed 538,000 entries
  2. UniProtKB/TrEMBL Redundant, automatically annotated – unreviewed 24,000,000 entries

www.uniprot.org
Manual annotation of UniProtKB/Swiss-Prot

**Sequence**

Splice variants

**Features**

UniProtKB

**Annotations**

Nomenclature:
- Protein names
- Gene names

Protein quaking
- Also known as:
  - Name: QKI
  - Synonyms: Qk, Qk1, Qka1

**Ontologies**

- Cell cycle
- DNA damage
- DNA repair
- Fatty acid biosynthesis
- Lipid synthesis
- Nucleus
- Polymer
- Disease
- Repeat
- Zinc-finger
- DNA-binding
- Metal-binding
- Anti-oncogene
- Phosphorylation
- 3D-structure

**References**


Cited with:
- PMID: 12065031
- EMBL: Y123456
- GenBank: X234567

**Manual annotation of UniProtKB/Swiss-Prot**

- Isoform 1 (identifier: QKI/SS-1)
  - Also known as QKI-5,
  - Also known as QKI-7,
  - Also known as QKI-9
  - The sequence of this isoform differs from the canonical sequence as follows:
    - 312-341 GAVATKVRHMIRMPYQCRGTVCAAGGNGQG - WLYGQKSNRSVTEFSSDULITNA

- Isoform 2 (identifier: QKI/SS-2)
  - Also known as QKI-6
  - The sequence of this isoform differs from the canonical sequence as follows:
    - 312-341 GAVATKVRHMIRMPYQCRGTVCAAGGNGQG - EGMHPIFQPDIAH
Complete Proteomes

A proteome consists of the complete set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

The taxonomy pages provide links to download Complete Proteome Sets when available, as well as links to the HAMAP and/or IntegR8 web sites.

Search completely sequenced organisms:
Downloads

UniProt is updated every four weeks (see FAQ on how to be notified automatically of updates). You can download small data sets and subsets directly from this website by following the download link on any search result page. For downloading complete data sets we recommend using ftp.uniprot.org. If you are located in Europe, the Middle East or Africa, you may want to download data from our mirror site in the United Kingdom or in Switzerland instead.

Here are some direct links to frequently downloaded files:

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
<th>fmt</th>
<th>diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProtKB</td>
<td>UniProtKB/Swiss-Prot</td>
<td>xml</td>
<td>fasta</td>
</tr>
<tr>
<td></td>
<td>UniProtKB/EMBL</td>
<td>fasta</td>
<td>text</td>
</tr>
<tr>
<td></td>
<td>Isoform sequences</td>
<td>fasta</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Taxonomic divisions</td>
<td>text</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="#">Proteomes</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Only a selected set of species are available on the FTP site. All species can</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>be downloaded from the website.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ID mapping data</td>
<td></td>
<td>tab</td>
</tr>
<tr>
<td></td>
<td>Documents</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XML Schema</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniRef</td>
<td>UniRef00</td>
<td>xml</td>
<td>fasta</td>
</tr>
<tr>
<td></td>
<td>UniRef90</td>
<td>xml</td>
<td>fasta</td>
</tr>
<tr>
<td></td>
<td>UniRef80</td>
<td>xml</td>
<td>fasta</td>
</tr>
<tr>
<td></td>
<td>XML Schema</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniParc</td>
<td>UniParc (Sequence Archive)</td>
<td>xml</td>
<td>fasta</td>
</tr>
<tr>
<td></td>
<td>XML Schema</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniMES</td>
<td>Metagenomic and Environmental Sequences</td>
<td>fasta</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UniMES matches to InterPro methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniProt RDF</td>
<td>UniProt RDF distribution</td>
<td></td>
<td>rdf</td>
</tr>
</tbody>
</table>
Reference Proteomes

Complete proteome – a set of proteins thought to be expressed by organisms whose genomes have been completely sequenced.

For an increasing number of organisms, more than one genome have been sequenced at the species level i.e. multiple strains.

Literature functional studies, strain is often not identified
Reference Proteomes

- UniProt now identifies Reference Proteomes - a set of protein sequences derived from a complete proteome which constitutes a defined standard for a particular user community.
- Reference proteomes are manually defined – generally the proteomes of well-studied model organisms and other proteomes of interest for biomedical and biotechnological research.
- Reference proteomes have been selected to provide broad coverage of the tree of life, and constitute a representative cross-section of the taxonomic diversity to be found within UniProtKB.
UniProt and ‘Omics’ data

- UniProt database of choice of proteomics/interactomics community

Interactions can be mapped to the canonical sequence...

.. to splice variants...

.. or to post-processed chains
PRIDE – MS identifications

• Stores submitted MS identifications plus meta-data describing sample
• PRIDE-Q will hold a filtered subset- high quality identifications
• Will then build an expression atlas to give organism/cellular/subcellular map at the protein level
• Sample annotation enables comparison of normal/cell line/diseased cell protein profile
• Protein identification provides proof of protein existence and will also be a source of PTM data.
PRoteomics IDEntifications database (PRIDE)

Introduction

The PRIDE PRoteomics IDEntifications database is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence.

Submitting Data to PRIDE

PRIDE encourages and welcomes direct user submissions of mass spectrometry data to be published in peer-reviewed publications.

Search PRIDE

Examples:

- PRIDE Experiment accession number: e.g. 8500
- UniProtKB Ac: e.g. P29375
- UniProtKB Id: e.g. KDM5A_HUMAN
- IPI: e.g. IPI00619935
- Ensembl: e.g. ENSP0000038288

You can also browse PRIDE by species, tissue, cell type, GO terms and disease.

More complex searches can be performed using the Advanced Search.

You can use the PRIDE BioMart for custom queries, such as linking PRIDE identifications to REACTOME pathways.

PRIDE Statistics

- 26,021 Experiments
- 11,206,694 Identified Proteins
- 62,037,413 Identified Peptides

News

PRIDE at EBI

pride_ebi

pride_ebi dataset by M Nemoto:

PRIDE Inspector

Key features: [More]

- Visualize MS data in PRIDE XML and mzML format
Meta-data is indexed using controlled vocabulary terms
‘Omics analysis

• The end result of such experiments is usually a list of identifiers

• Biological context needs to be derived from this list

• Tools available for this, using EBI data
  • Network analysis
  • Pathway analysis
  • GO representation analysis
Network analysis

- Protein interaction networks aim to represent the complex web of molecular interactions within a cell.

- Overlay of expression data on those networks can help identify clusters of interacting (i.e. probably functionally related) proteins which are up(down)-regulated in your sample.

- Graph analysis tools (e.g. as supplied in R) can enable a deeper understanding of the data.
IntAct

• Publicly available repository of molecular interactions (mainly PPIs) - ~305K binary interactions taken from >5,600 publications (October 2012)

• Data is standards-compliant and available via our website, for download at our ftp site or via PSICQUIC

  http://www.ebi.ac.uk/intact
  ftp://ftp.ebi.ac.uk/pub/databases/intact
  www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml

• Provide open-access versions of the software to allow installation of local IntAct nodes.
IntAct – Home Page

IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. To perform a search in the IntAct database use the search box in the top left corner.

Examples:
- Gene name: e.g. BRCA2
- UniProtKB Ac: e.g. Q05609
- UniProtKB Id: e.g. dmc1
- Pubmed Id: e.g. 10831611

Please supply your feedback to helpdesk. We thank you for your help in further developing IntAct.

Citing IntAct
- The IntAct molecular interaction database in

Dataset of the month: May
- Systematic analysis of dimeric E3-RING Interactions reveals increased combinatorial
Use IntAct to….

• Search for the interactors of one, or a set of, proteins

• Perform an ontology-driven search (e.g. find all the interaction of human peroxisomal proteins – GO:0005777 peroxisome)

• Download a large dataset (e.g. the human interactome) in standard formats (XML, tab-delineated, XGMML, RDF)
Network analysis – adding quantitative expression data to an interaction network
Gene ontology enrichment analysis

Choose Layouts → Cytoscape Layouts → Hierarchical layout

Child terms
- negative regulation of transcription from RNA polymerase II promoter
- regulation of macromolecule metabolic process

Parent terms
Pathway analysis of large datasets

• Network analysis highlights interacting clusters of molecules (often indicates protein complexes stable/transient) – may have similar function and subcellular location

• To group molecules by Process, can also subject identifier link to pathway analysis
A Database of human biological pathways
“Caspase-8 is the key initiator caspase in the death-receptor pathway. Upon ligand binding, death receptors such as CD95 (Apo-1/Fas) aggregate and form membrane-bound signalling complexes (Box 3). These complexes then recruit, through adapter proteins, several molecules of procaspase-8, resulting in a high local concentration of zymogen. The induced proximity model posits that under these crowded conditions, the low intrinsic protease activity of procaspase-8 (ref. 20) is sufficient to allow the various proenzyme molecules to mutually cleave and activate each other (Box 2). A similar mechanism of action has been proposed to mediate the activation of several other caspases, including caspase-2 and the nematode caspase CED-3 (ref. 21).”

How can I access the pathway described here and reuse it?
Rationale - Figures

A picture paints a thousand words...

but....
• Just pixels
• Omits key details
• Assumes
• Fact or Hypothesis?
Reactome is...

Free, online, open-source curated database of pathways and reactions in human biology

Authored by expert biologists, maintained by Reactome editorial staff (curators)

Mapped to cellular compartment – agnostic of cell/tissue type/state
Data Expansion - Link-outs From Reactome

• GO
  • Molecular Function
  • Compartment
  • Biological process

• KEGG, ChEBI – small molecules
• UniProt – proteins
• Sequence dbs – Ensembl, OMIM, Entrez Gene, RefSeq, HapMap, UCSC, KEGG Gene
• PubMed references – literature evidence for events
Species Selection

Switch Species: Homo sapiens

Search results:
- Saccharomyces cerevisiae
- Schizosaccharomyces pombe
- Canis familiaris
- Rattus norvegicus
- Bos taurus
- Drosophila melanogaster
- Plasmodium falciparum
- Staphylococcus aureus N315
- Homo sapiens
- Taeniopygia guttata
- Sus scrofa
- Dictyostelium discoideum
- Arabidopsis thaliana
- Mus musculus
- Mycobacterium tuberculosis
- Gallus gallus
- Xenopus tropicalis
- Oryza sativa
- Caenorhabditis elegans
- Escherichia coli
Data Expansion – Projecting to Other Species

Human

\[ A + ATP \rightarrow A - P + ADP \]

Mouse

\[ A + ATP \rightarrow A - P + ADP \]

Drosophila

\[ \text{Reaction not inferred} \]

No orthologue - Protein not inferred
Reactome Tools

• Interactive Pathway Browser

• Pathway Mapping and Over-representation

• Expression overlay onto pathways

• Molecular Interaction overlay

• Biomart
Pathway Analysis

Allows you to analyse a list of protein, gene, expression data or compound identifiers and determine how they are likely to affect pathways. More...

Paste or upload your data:

- 000139
- 000186
- 000187
- 000204
- 000217
- 000231
- 000232
- 000233
- 000254
- 000267
- 000268
- 000273
- 000303

Select your desired analysis tool

Inhouse services:

- ID mapping and pathway assignment. Takes your list of IDs and finds the corresponding pathways from Reactome, plus the corresponding Uniprot IDs.

- Overrepresentation analysis. Finds the Reactome pathways in which IDs in your list are strongly enriched - can help to understand the biological context of your data.

EMBL-EBI
Pathway Analysis – Overrepresentation

Each event is coloured according to the un-adjusted, i.e., not corrected for multiple testing, probability (from hypergeometric test) of seeing given number or more genes in this event by chance. Please note that only those ‘child’ events are shown which have a p-value lower than the ‘parent’ event. The top-level (root) events are ordered according to the lowest p-value of their components.

Colour key for probabilities:
1e+00, 3e+00, 6e+00, 9e+00, 1e+01, 3e+01, 6e+01, 9e+01, 1e+02, 3e+02, 6e+02, 9e+02, 1e+03, 3e+03, 6e+03, 9e+03, 1e+04, 3e+04, 6e+04, 9e+04, 1e+05, 3e+05, 6e+05, 9e+05

- Cell Cycle, Mitosis
- Gene Expression
- DNA Replication
- Signalling by NGF
- Apoptosis
- DNA Repair
- Transcription
- Cell Cycle Checkpoints
- Metabolism of carbohydrates
- Integrin cell surface interaction
- RNA Processing
- HIV Infection
- Apoptosis
- Chromosome Maintenance
- Signalling by Wnt
- CoCo2Pho: APO/C mediated degradation of Cyclin
- Regulation of activated PAK2 by proteasome mediated degradation
- Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins
- Signalling in immune system
- Signalling by EGFR
- Transmembrane transport of small molecules
- SLC-mediated transmembrane transport
- Urea cycle
- Glycolysis and Gluconeogenesis
- Metabolic network

P-val

Reveal next level

‘Top-level’
Expression Analysis I

Upload expression data

Takes gene expression data (and also numerical proteomics or metabolomics data) and shows how expression levels affect reactions and pathways in living organisms. May be time-consuming, depending on the number of identifiers you are submitting: less than 5000: a few seconds, 5000 - 10000: a few minutes, 10000 or more: 10 minutes or longer. More....

Paste or upload your data:

```
#Probeset  10h_control  10h 14h 18h 24h
1053_at  8.040078  7.147358  6.706705  6.794622  7.475157
1729_at  6.895988  6.991047  12.9922  7.112222  7.04721
200002_at  12.555275  12.511045  12.564419  12.538642  12.439174
200003_s_at  12.401259  12.054083  12.275169  12.206342  12.015475
200012_x_at  12.486269  12.402275  12.302666  12.256543  12.232444
200016_x_at  12.110458  11.93288  11.938524  11.89243  11.468105
200022_at  12.205038  11.927471  12.064725  12.031422  11.932255
```

[Example] [Clear]

Browse  Analyse
Expression Analysis II

'Hot' = high
'Cold' = low

Step through Data columns
PST

• Databases exist to hold proteomics data - now need to find
  • Creative ways to display the information using data from different sources e.g. the human hepatocyte interactome
  • Easier ways for wet-lab scientists to analyse their data, less dependency on Cytoscape for initial work
  • Work flows for ‘Omics data analysis