Proteomics: An introduction to EMBL-EBI resources

Sandra Orchard [1]
Proteins
Beginner
0.5 hour

This course will provide an overview of proteomics. We will then introduce resources provided by the EBI related to proteomics.

Learning objectives:

- Understanding what is meant by 'proteomics'
- Good understanding of which proteomics resources are available at the EBI
- Understanding of what the UniProtKB, IntAct, Reactome and PRIDE databases are
- Knowing what the differences between the UniProtKB, IntAct, Reactome and PRIDE databases are, and when to use each database

What is proteomics?

Proteomics [2] is the large-scale study of proteomes. A proteome [3] is a set of proteins produced in an organism, system, or biological context. We may refer to, for instance, the proteome of a species (for example, Homo sapiens) or an organ (for example, the liver). The proteome is not constant; it differs from cell to cell and changes over time. To some degree, the proteome reflects the underlying transcriptome [4]. However, protein activity (often assessed by the reaction rate of the processes in which the protein is involved) is also modulated by many factors in addition to the expression level of the relevant gene.

Proteomics is used to investigate (see Figure 1):

- when and where proteins are expressed;
- rates of protein production, degradation, and steady-state abundance;
- how proteins are modified (for example, post-translational modifications (PTMs) such as phosphorylation);
- the movement of proteins between subcellular [5] compartments;
- the involvement of proteins in metabolic pathways;
- how proteins interact with one another.

Proteomics can provide significant biological information for many biological problems, such as:

- Which proteins interact with a particular protein of interest (for example, the tumour suppressor protein p53)? (Human example [6])
- Which proteins are localised to a subcellular compartment (for example, the mitochondrion)? (Human example [7])
- Which proteins are involved in a biological process (for example, circadian rhythm)? (Human example [8])

Several high-throughput technologies [9] have been developed to investigate proteomes in depth. The most commonly applied are mass spectrometry [10] (MS)-based techniques such as Tandem-MS [11] and gel-based techniques such as differential in-gel electrophoresis (DIGE [12]). These high-throughput technologies generate huge amounts of data. Databases are critical for recording and
carefully storing this data, allowing the researcher to make connections between their results and existing knowledge. The EBI [13] hosts up-to-date and accurate databases to enable rapid searching and retrieval of these data. The four major databases related to proteomic research (UniProtKB [14], IntAct [15], Reactome [16] and PRIDE [17]) are described in the next section. These four databases (especially UniProtKB) draw from gene sequence data (e.g. Ensembl [18]) and annotation [19] tools (e.g. InterPro [20]) also hosted by the EBI. You can find out more about such resources in other training courses, such as the Introduction to Functional Genomics [21] Resources [21], and the InterPro Quick Tour [22].

**Figure 1. Areas of proteomics.** Proteomic experiments generally collect data on three properties of proteins in a sample: location, abundance/turnover and post-translational modifications. Depending on the experimental design, researchers may be directly interested in these data, or may use them to infer additional information. For example, it may be possible to infer a protein’s interaction partners among others that are colocalised with it, or to assess whether a protein is active from its post-translational modifications.

**Proteomics resources at the EBI**

This section will introduce you to the proteomics [2] resources at the EBI: UniProtKB [14], IntAct [15], Reactome [23] and PRIDE [17] (Figure 2):
the **UniProt Knowledgebase** [14] (UniProtKB [24]) database contains protein sequences and information about the known biological functions (e.g. 'an ATP synthase') of proteins (1 [25]);

the **IntAct** [15] database contains information about protein interactions (2 [25]);

the **Reactome** [16] database contains information about which roles proteins play in human biological pathways, and which processes they contribute to (3 [25]);

the **PRIDE** [17] database contains experimental evidence of published protein and peptide identifications (4 [25]).
Experimental evidence can be analysed with UniProtKB to generate new protein identifications. The raw data [26] is uploaded to PRIDE in support of the published results. Curated results demonstrating protein-protein interactions or evidence of association with a biological pathway are added to IntAct and Reactome respectively. All four databases use UniProt accession [27] numbers as a standard method of referencing proteins.

Figure 2. Association of data between UniProtKB, IntAct, Reactome and PRIDE. Experimental evidence can be analysed with UniProtKB to generate new protein identifications. The raw data [26] is uploaded to PRIDE in support of the published results. Curated results demonstrating protein-protein interactions or evidence of association with a biological pathway are added to IntAct and Reactome respectively. All four databases use UniProt accession [27] numbers as a standard method of referencing proteins.

UniProtKB

UniProtKB [28] is a database of protein sequences and functional information ([1] [29]). It is part of the UniProt [14] resource, which also includes databases of clustered sequences (UniRef [30]) and metagenomic data (UniMES [31]), and an archive of sequence data that is in the public domain [32] (UniParc [33]). UniProtKB consists of two sections:

- UniProtKB/Swiss-Prot [34]. The information in this section of the database is manually annotated [35] and reviewed, so it is of high quality and is non-redundant.
UniProtKB/TrEMBL [36]. The information in this section is computationally annotated and not reviewed, so it provides high annotation coverage of the proteome [37].

The majority of sequences in UniProtKB (~85%) are originally derived from translations of genetic coding sequences submitted to the public nucleic acid databases, the ENA [38]/GenBank [39]/DDBJ [40] databases. Translated sequences are automatically added to UniProtKB/TrEMBL, and are migrated to UniProtKB/SwissProt after manual curation [41].

You can use UniProtKB to find a wealth of information on a protein of interest. For example, you can find evidence for the structure or function of a protein, summarised from peer-reviewed papers, and evidence for subcellular [42] location or involvement in disease. UniProtKB also enables you to compare protein sequences to investigate areas of homology.

IntAct

IntAct [15] is a database of molecular interactions, in particular protein-protein interactions. The data in IntAct are obtained from the scientific literature or from direct data submissions by expert curators following a detailed annotation [19] model (2 [29]). The annotation of these interaction data complies with international standards agreed between the majority of protein interaction databases (the IMEx [43] Consortium).

You can search IntAct for the interaction partners of: a specific protein; a set of proteins; and/or a particular organism. IntAct provides the protein interactions together with information on which experimental methods were used to detect the interaction and a reference to the publication in which the interactions are described. Proteins in IntAct are mapped to UniProtKB [44] accession numbers, so you can link from protein entries in IntAct straight to the UniProtKB summary page. Similarly, small molecules [45] are linked to the corresponding ChEBI [46] entry.

Reactome

Reactome [23] is a human-centric database of the roles that proteins play in biological pathways. While IntAct [47] focuses on the interaction of proteins with other target molecules, Reactome looks at proteins associated with biological processes, breaking each process down into a series of molecular events. Reactome events have inputs (substrates) and outputs (products). Like IntAct, Reactome is curated by life scientists, using evidence from the scientific literature and input from experts in each pathway. The curation [41] rules can be found on the Reactome wiki [48].

The Reactome Pathway Analysis Tool can be used to 'paint' pathways with a list of proteins (e.g. a list of UniProt [49] accessions derived from a proteomics [2] experiment) or genes (e.g. a list of EMBL/GenBank/DDBJ IDs derived from a functional genomics [50] experiment). Reactome will display pathways associated with your submitted data, with links from within Reactome to biological process entries. Reactome uses UniProtKB [44] IDs, which can be followed to a UniProtKB summary page.

PRIDE

PRIDE [17] is a database of Proteomics [2] IDEntifications. It contains protein and peptide identifications (including details of post-translational modifications [51]) together with the mass
spectrometric evidence supporting these identifications. PRIDE acts as a repository for mass spectrometry data, specifically the fragment ion spectra used to identify peptide sequences (4 [29]). You can follow protein identifications back to their UniProtKB [44] page using the UniProtKB accessions.

You can use PRIDE to find data that have been submitted in support of a piece of published work, or you can search across submissions for particular proteins of interest. Each identification carries with it a description of the experimental method used. PRIDE links the published work from the summary page using PubMed [54] IDs. An increasing number of journals mandate deposition of data in PRIDE as part of the publication process, hence the quality of data in PRIDE broadly follows the protein identification standards required by journals. Most journals require the minimum level evidence specified by the 'Philadelphia Guidelines' (6 [29]) and fully described in the MIAPE guidelines. You can learn more about PRIDE on the PRIDE Quick Tour [55].

Summary

Proteomics [2] is the large-scale study of proteomes. A proteome [3] is a set of proteins produced in an organism, system, or biological context.

Major proteomics resources at the EBI include UniProtKB [44], IntAct [47], Reactome [56] and PRIDE [57]:

- the UniProt Knowledgebase [14] (UniProtKB [24]) contains protein sequences and functional information (1 [25]);
- the Reactome [16] database contains information about protein involvement in human biological pathways (3 [25]);
- the PRIDE [17] database contains experimental evidence of published protein and peptide identifications (4 [25]).

UniProtKB can be used to find the sequence of a protein of interest, and to see what functional information has been ascribed to it. You can then use IntAct to look for published evidence of interactions involving your protein, or look for involvement in known biological processes with Reactome. Finally, you can search PRIDE to find the data from experiments where your protein was identified.

Quiz: Proteomics at the EBI

Questions: 6
Attempts allowed: Unlimited
Available: Always
Pass rate: 75 %
Backwards navigation: Allowed

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.
References


Learn more

Find out more

You can find more in-depth detail about the UniProt [49] databases in the UniProt Quick Tour [64] or the help page [65].

There are also Quick Tours available for IntAct [66] and PRIDE [55].

The EBI offers hands-on courses covering a range of subjects, including training for Proteomics [2] resources. An up-to-date list of training courses is available here [67].

Contributors

Sandra Orchard [68]
EMBL-EBI
Proteomics Services Team Coordinator

Sandra Orchard leads the Molecular Interaction Team, responsible for developing resources enabling...
the network analysis of large-scale datasets as well as supplying basic interaction data and information about protein complexes. She is responsible for the production and maintenance of the IntAct Molecular Interaction database and the Complex Portal. She has previously contributed to the annotation of the UniProtKB, InterPro and GOA databases. She also applies her experience to provide hands-on training in several resources, including UniProtKB, InterPro, IntAct and Reactome.

Phil Charles [69]

University of Cambridge/EBI

Phil is a Ph.D student in Kathryn Lilley's lab at the Cambridge Centre for Proteomics, in the University of Cambridge Department of Biochemistry.

All course materials in Train online are free cultural works licensed under a Creative Commons Attribution-ShareAlike 4.0 International license

Source URL: http://www.ebi.ac.uk/training/online/course/proteomics-introduction-ebi-resources

Links
[1] http://www.ebi.ac.uk/training/online/trainers/orchard%40ebi.ac.uk
[10] http://www.ebi.ac.uk/training/online/glossary/mass-spectrometry
[12] http://www.ebi.ac.uk/training/online/glossary/dige
[17] http://www.ebi.ac.uk/pride/