Protein interactions and their importance

Pablo Porras Millan [1]

- Systems
- Proteins
- Beginner
- 1 hour

This course will provide an introduction to molecular interactions, their importance and the methodologies use to generate and capture interaction data.

Learning objectives:

- Understanding the importance of molecular interaction information and be able to provide examples of different types of molecular interaction
- Be able to list the main experimental methodologies used to study protein-protein interactions
- Be aware of the limitations that these methodologies have
- Knowing where to find molecular interaction data at EMBL-EBI

Protein-protein Interactions

Understanding physical and functional interactions between molecules in living systems is of vital importance in biology. Several powerful methodologies and techniques have been developed to generate molecular interaction data, concentrating mainly on protein–protein interactions (Figure 1) [2].

Given the importance of protein-protein interactions and their vast numbers in comparison with datasets involving other types of molecules, we focus on them in this course.

Molecular interaction data can be generated using many different techniques, all of which have their strengths and weaknesses. However, it is important to stress that all molecular interaction data is to some degree artifactual. No single method can accurately reproduce a true binary interaction observed under physiological conditions.

The "boom" in molecular interaction research that we have experienced in the past few years has been caused by the increasingly wide availability of high throughput technologies that can potentially provide information on several thousand pairwise interactions at a time [1]. Such high-throughput studies can provide a global 'snapshot' of the molecular interactions that take place in a cell, an organism or as particular physiological context. This is known as the interactome [3].

Understanding the cellular machinery and identifying interactions that underpin particular physiological processes relies on the retrieval, organisation and analysis of these valuable data [1]. Efforts have been made in the protein interaction field towards addressing this challenge.
This course will review some of the main techniques used to produce protein interaction data and discuss their respective advantages and disadvantages. We will discuss how you should regard the reliability of each of the methods. We will also explain how the experimental data are captured electronically.

In nucleotide sequence databases, sequence data is represented simply as a string of letters. Representing protein interaction data is somewhat more complex. We need to use the correct identifiers for the molecules reported to interact; we also need to record the method used to detect the interaction, among other relevant information. Although there are ongoing significant international efforts to standardise how such information is reported and described and enable exchange of data among different public repositories, this field is not as mature as the nucleotide sequence-data field.

The importance of molecular interactions

Molecular interactions are important to molecular biologists because:

1. They help us to understand a protein's function and behaviour (Figure 2).

2. They can help us to predict the biological processes that a protein of unknown function is involved in:
   
   - We may assume “Guilt by association” if a protein of unknown function associates with one of known function
   - Proteins involved in the same process should cluster together in network maps

3. They can help us to characterise protein complexes and pathways; interaction networks can be used as a draft 'map' to add detail to biological processes and pathways.
Contrary to the original belief that one protein had a single function, proteins have different functions and cellular roles depending on their immediate environment, which has an impact on their position in protein networks.

Where do the data come from?

Methods for molecular interaction identification

There are two approaches to gaining information about molecular interactions:

- computational
- experimental

In this course we will concentrate on the experimental methods, but there is an increasing variety of computational methods that can predict protein–protein interactions. Because only a small proportion of all the molecular interactions in an organism are currently covered by the experimental data, these methods provide a meaningful resource that can help use to analyse under-represented regions of the interactome [3]. You can find out more about the different methods used to predict protein-protein interactions in a comprehensive Wikipedia entry entitled protein-protein interaction prediction [4].

A wide variety of experimental methods can be used to detect protein-protein interactions. It is important to realise that there is no perfect approach. Each method has its limitations and is to an extent potentially artifactual. Therefore, it is advisable to check interactions using more than one approach: interactions detected by more than one method are more likely to be "real" [2].

Next, we will have a look at some ways to experimentally identify protein-protein interactions.

High-Throughput: Yeast two hybrid
The most frequently used laboratory method for experimentally determining molecular interactions is **yeast two-hybrid (Y2H)** screening\(^2\) [2].

Y2H is a complementation **assay** [5]. The readout mechanism is based on a transcription factor, which is split into two independent parts, the DNA-binding **domain** [6] (BD) and the DNA-activation domain (AD). The BD and AD domains are fused to two proteins of interest, the bait (X) and the prey (Y). This ensures that the readout can only take place when the two halves are brought into close proximity. If the bait and prey proteins bind to each other when expressed in a yeast cell, the transcription machinery becomes activated and a reporter gene is turned on (Figure 3).

**Figure 3.** The yeast two hybrid (Y2H) concept and a typical readout. [1] The BD domain fused to the bait protein (X) and the AD domain fused to prey protein (Y) are expressed in yeast cells. [2] If proteins X and Y interact, BD binds DNA and AD activates RNA polymerase.

An example readout [3] of a Y2H assay with two bait proteins (Bait 1 and Bait 2) and five prey proteins (1 to 5). In this example, positive interactions are shown by colony growth. Readouts for a Y2H assay can also be detected by DNA sequencing or colorimetric methods such as the beta-galactosidase assay (Figure from Koh et. al.\(^5\) [2]).

**Advantages:**

- Fast
- Inexpensive
- Scalable
- An *in vivo* system in which binding sites can be accurately mapped
Disadvantages:

- False positives occur when a yeast protein acts as a bridge for the interaction.
- Interactions occur between proteins that would not normally be present in the same cellular compartment, in the same cell type, or at the same time.
- Both bait and prey proteins can fail to be expressed or might be toxic to the cell.

High-Throughput: Affinity Purification Mass Spectrometry

The second high-throughput method significantly contributing to the growth in published protein-protein interaction data is affinity purification mass spectrometry (AP-MS, Figure 4).


In AP-MS, a single protein or molecule of interest is immobilised in a matrix as a bait. Then a protein mixture is passed through the matrix and interacting partners (prey) are captured by the bait protein. Any form of technique relying on mass spectrometry [9] (MALDI, LC-MS/MS, etc...) is then used to identify the captured proteins.

The affinity purification step and type of mass-spectrometry-based identification can be modified. For example, you can perform an immunoprecipitation and then identify the captured proteins using LC-MS/MS or you can perform a pull-down of epitope-tagged molecules and then identify the proteins using MALDI-MS [2].

Figure 4. Affinity purification and mass spectrometry (AP-MS). The bait protein (yellow) is immobilised on a matrix [1]. A protein mixture is passed through and only the interacting partners (prey) are retained [2]. In the following step the prey proteins are removed, digested with a protease and the resulting peptides are analysed by MS [3] [2].
Advantages:

- Potentially, depending on the sensitivity of your MS-approach and the affinity of the interacting partners, this method has the ability to examine interactions among multiple proteins at subpicomole concentrations.
- The prey proteins are present in their native state (so long as they are not affected by the sample lysis process) and concentration.

Disadvantages:

- Prey proteins without a peptide signature recognisable by MS (owing to obscure post-translational modifications) or present in very low amounts will not be identified.
- Biologically relevant transient interactions and weak interactions may be missed.
- Mixing of compartments during cell lysis/purification is a potential source of false positives. For example, interactions between proteins that would not normally be in the same cellular compartment may confound your results.

There is an overwhelming variety of techniques that can be used to detect protein-protein interactions using low- or medium-throughput setups. Next we summarise some methods that are often used to improve the confidence of an interaction detected by high-throughput methods or on their own merits in small-scale experiments.

Co-immunoprecipitation

Co-immunoprecipitation (Co-IP) is the immunoprecipitation of intact protein complexes (i.e. antigen along with any proteins or ligands that are bound to it); see Figure 5. Co-IP works by selecting an antibody that targets a known protein that is believed to be a member of a larger complex of proteins. By targeting this known member of a complex with an antibody, you might be able to pull the entire protein complex out of solution and thereby identify unknown members of the complex.

This technique works when the proteins involved in the complex bind to each other tightly, making it possible to pull multiple members of the complex out of solution by latching onto one member with an antibody.

The concept of pulling protein complexes out of solution is sometimes referred to as a "pull-down".

Co-IP has been traditionally considered as the "gold standard" assay [5] for protein-protein interactions, but its caveats are very similar to those of AP-MS [10] as it is also an affinity purification method.
Protein interactions and their importance
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)


X-ray crystallography

X-ray crystallography [12] is a method of determining the arrangement of atoms within a crystal, in which a beam of X-rays strikes a crystal and causes the beam of light to spread into many specific directions. From the angles and intensities of these diffracted beams, a crystallographer can produce a three-dimensional picture of the density of electrons within the crystal (Figure 6). From this electron density, the mean positions of the atoms in the crystal can be determined, as well as their chemical bonds and other types of information.

This method is considered to be another "gold standard" because it provides an extremely deep level of detail about interacting surfaces and residues (at a level of atoms and chemical bonds) and high quality data.

However, it is extremely challenging technically, is very low-throughput and is not free from false negatives or false positives; not every protein is amenable to co-crystallization and some proteins that co-crystallise in vitro do not interact in a physiological context.

Figure 6. X-ray crystallography is used to to obtain detailed structural and chemical insights for selected interactions. The figure shows a model of the cullin complex [2].

Other methods

Sometimes it is necessary to use methods that can be performed in mammalian cell lines, providing a more...
physiological environment for studies using mammalian proteins. The following techniques can be applied in medium- or high-throughput setups and have become widely used in the past few years:

**LUMIER [13]:** luminescence-based mammalian interactome [3] mapping

**MAPPIT [14]:** mammalian protein-protein interaction trap

**FRET [15]/BRET [16]:** fluorescence/bioluminescence-resonance energy transfer

For more information on these techniques, see Reference 9 [2].

Finally, one of the few ways of identifying transient interactions missed by other methods is the enzyme assay [5]. These assays are based on taking enzyme-catalysed reactions as evidence that an enzyme interacts with its substrate, for example. However, these assays can only use *in vitro* data, requiring purified proteins, as there are too many unknowns if they are performed with a whole cell lysates. Moreover, many enzymes are promiscuous *in vitro* – most prominently kinases. This can lead to a large number of false positives.

### Interaction databases

Given the variety of techniques available to produce protein-protein interaction data and the large number of studies that are published every day, an enormous effort is required to store this information in a way that is both accessible and intelligible to the user (Figure 7).

Molecular interaction databases aim to fulfill this need by extracting information from scientific publications or, in some cases, by including protein-protein interaction predictions found using computational methods 14 [2].

The storage of interactions in publicly available databases allows access to a large volume of interaction data and subsequent analysis of the interactome [3].
Protein interactions and their importance
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

Figure 7. Database storage and analysis of interaction data. Data obtained by interaction detection methods is stored in databases and available for analysis; here we show a molecular interaction network as one means of visualising such an analysis. Figure adapted from Koh et al. 5

There are several issues when representing interaction data, including use of nomenclature, representing complex data and cross-referencing to other resources. Different databases have different approaches to solving these issues, and database teams are constantly developing new strategies to improve their representation of the data. They also have different levels of curation 17, depending on how much detail they capture about each interaction.

We will now discuss how heterogeneous database representation of interactions is, and describe some of the initiatives that aim to reduce this heterogeneity.

Different types of molecular-interaction databases

There are a large number of publicly available molecular interaction databases. These can be classified into three main types, according to their data-acquisition policies 14 2.

Primary databases are those that collect experimental molecular interaction data exclusively from peer-reviewed scientific publications. IntAct 18 11 2, MINT 19 2 and MatrixDB 16 2 are examples of this type of database. They can be further classified by the level of detail that they use to represent the information and the depth of their curation 17 policies.
Secondary databases, also known as meta-databases, seek to integrate the data curated by several primary databases in one, integrated repository. APID \(^{17}\) [2] and PINA \(^{18}\) [2] are examples of this type.

Finally, predictive databases combine the experimentally inferred data taken from primary databases with computational predictions of molecular interactions. Examples include STRING \(^{19}\) [2] and UniHI \(^{20}\) [2].

Now, let's have a look at the different types of curation that primary databases use to register molecular interaction data.

### Curation levels

As explained before, different databases use varying levels of annotation \(^{19}\) or curation \(^{17}\). Figure 8 organises some molecular interactions and pathways databases, including members of the IMEx consortium \(^{20}\), according to the different levels of detail in their curation procedures, ranging from light curation to the very detailed level of description that the IMEx guidelines \(^{21}\) require.

<table>
<thead>
<tr>
<th>DEEP CURATION</th>
<th>SHALLOW CURATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>MatrixDB</td>
<td>BioGRID</td>
</tr>
<tr>
<td>IntAct</td>
<td>InnateDB</td>
</tr>
<tr>
<td>MINT</td>
<td>Human Protein Reference Database</td>
</tr>
<tr>
<td>MPact</td>
<td>InnateDB</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>InnateDB</td>
</tr>
<tr>
<td>J. Craig Venter</td>
<td>InnateDB</td>
</tr>
<tr>
<td>STRING</td>
<td>InnateDB</td>
</tr>
<tr>
<td>VIR2L</td>
<td>InnateDB</td>
</tr>
<tr>
<td>HOSTNET</td>
<td>InnateDB</td>
</tr>
<tr>
<td>J. Craig Venter</td>
<td>InnateDB</td>
</tr>
<tr>
<td>INSTITUTE</td>
<td>InnateDB</td>
</tr>
</tbody>
</table>

**Figure 8.** Some molecular interactions and pathways databases organised according to their level of curation.
Light curation is used where the main goal is to re-publish maximum content with minimum effort - sufficient to give identifiers to the interacting partners and state the interaction detection method used. The MIMIx (Minimum Information required for reporting a Molecular Interaction eXperiment) standard goes one step beyond this, providing a lightweight version of the IMEx guidelines [12,15] [2].

Deep curation requires a detailed description of all the features involved in the interaction and interacting partners, and complies with the full version of the IMEx guidelines; this requires more time and resources. IMEx members including IntAct [18], MINT, DIP and MatrixDB are required to register their data complying with these guidelines. You can see a full list of IMEx-compliant databases on the IMEx Consortium [22] website.

Now, let's have a look at the molecular interactions database hosted in EMBL-EBI.

Molecular interaction databases at EMBL-EBI

IntAct [23] is a central, public repository where molecular interactions data can be stored and accessed. It is hosted by the European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK, where it is maintained by a group of curators and developers.

We populate IntAct [24] with interaction data from literature curation [17] or direct user submissions. Most of the data refer to protein-protein interactions, but interactions involving other types of molecules, such as small chemical compounds or nucleic acids, can also be found in IntAct.

At EMBL-EBI IntAct is the main database for molecular interactions. There are other EMBL-EBI databases that also capture interaction information, the most prominent one being ChEMBL [25], which hosts a large collection of small molecule-protein/drug-target interactions.

IntAct is a member of the International Molecular Exchange (IMEx [22]) Consortium - a group of major public interaction data providers whose goal is to share curation effort and exchange completed records on molecular interaction data. When you query data in IntAct you also access over 150 million interactions in a further 31 data resources via our PSICQUIC [26] (Proteomics Standard Initiative Common QUery InterfaCe) service or a consistently annotated, non-redundant, experimentally determined subset from the IMEx Consortium [27].

You can learn more about IntAct in a separate course: IntAct: Molecular Interactions at the EBI [28].

Summary

- Knowledge of molecular interactions helps us to assign new functions to molecules with unknown roles and to draw interaction maps that can improve our knowledge of biological pathways.
• The largest amount of interaction data available in the public domain [6] is for protein-protein interactions; a wide variety of experimental approaches can be used to detect these.

• There is no perfect detection method for molecular interactions. Each method has its limitations and is to an extent potentially artefactual. Therefore, it is advisable to take complementary approaches to gain confidence that a particular interaction exists in a physiological context.

• The two most frequently used high-throughput methods for detecting protein-protein interactions are yeast two-hybrid (Y2H) screening and affinity purification mass spectrometry (AP-MS). There is a great variety of other methods, generally used for smaller-scale setups.

• At EMBL-EBI, IntAct [24] is the main database storing molecular interactions. IntAct is a member of the International Molecular Exchange (IMEx [22]) Consortium, an effort to standardise curation [17] procedures for molecular interactions and combine the work of different databases.

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.

Learn more

Find out more


• **Molecular interaction databases.** Orchard, S. Proteomics. (2012) 12(10): 1656-1662, [PMID:22611057] [31] - A comprehensive review about the main molecular interaction databases that are available at present.


• **PSICQUIC and PSISCORE - accessing and scoring molecular interactions.** Aranda, B. et al. Nat. Methods. (2011). 8(7): 528-529, [PMID:21716279] [34] - The paper in which PSICQUIC, a query interface that allows access to data from multiple molecular interactions and pathways databases, is presented.
Recommended courses

**IntAct: Quick tour** [35]

**IntAct: Molecular Interactions at the EBI** [28] - online course

**Proteomics: An introduction to the EBI resources** [36] - online course

**Networks and Pathways for Biologists** [37] - held at EMBL-EBI

References


### Need some help?

For support, submission and related enquiries, email the intact [18]-help [at] ebi.ac.uk (IntAct help desk)

### Contributors

[1]
Protein interactions and their importance
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

Pablo Porras Millan [1]
EMBL-EBI
Scientific curator in IntAct

Pablo Porras got his PhD in 2006 in the University of Córdoba, Spain, having done research about transmembrane protein translocation and redox homeostasis. After that, he moved to Berlin to work in the Neuroproteomics group of the Max Delbrueck Center, getting involved in projects dealing with interactomics, neurodegenerative diseases and the ubiquitin-proteasome system. During this postdoc, he faced the problem of how to represent and analyze molecular interactions data. This experience proved to be of great value once he joined the EBI to work as a scientific curator in the molecular interactions database IntAct [58] in 2011.

Mindi Sehra [59]
EMBL-EBI
EMBL-EBI ALUMNI Scientific Training Officer - eLearning Content Developer

Mindi Sehra was the Scientific Training Officer (eLearning) for the Outreach and Training Team at EMBL-EBI. Mindi was responsible for expanding and consolidating the EBI’s range of online training materials and monitoring and maintaining the portal, which includes investigating ways to exploit electronic technologies. Mindi completed a Genetics Degree at Sheffield University before moving into genome analysis at the Wellcome Trust Sanger Institute. She completed a MSc in Medical Genetics and Immunology at Brunel University, her thesis on the Swine Leukocyte Antigen secured her a position in the Human and Vertebrate Annotation [60] and Analysis group as a computer biologist. She then joined the UniProt [61] team at the EBI as a protein curator [62] working on automatic and manual annotation [63].

Source URL: https://www.ebi.ac.uk/training/online/course/protein-interactions-and-their-importance

Links
[1] https://www.ebi.ac.uk/training/online/trainers/pporras
[2] https://www.ebi.ac.uk/training/online/course/protein-interactions-and-their-importance/references
[3] https://www.ebi.ac.uk/training/online/glossary/interactome
[5] https://www.ebi.ac.uk/training/online/glossary/assay
[6] https://www.ebi.ac.uk/training/online/glossary/domain
[7] https://www.ebi.ac.uk/training/online/glossary/specificity
[8] https://www.ebi.ac.uk/training/online/glossary/sensitivity
[10] https://www.ebi.ac.uk/training/online/glossary/term/118
[11] https://www.ebi.ac.uk/training/online/glossary/extract
[12] https://www.ebi.ac.uk/training/online/glossary/x-ray-crystallography
[16] http://en.wikipedia.org/wiki/F%C3%B6rster_resonance_energy_transfer#BRET