Report of Laboratory of Bioinformatics Alessandro Caula
Academic year 2019/2020

1. INTRODUCTION TO PROTEIN-LIGAND INTERACTION

Protein-protein interactions (PPIs) play a key role in any cellular process, the characterization and description of their behaviour is crucial for understanding the proteins role and function.
There are three major approaches used for PPIs studies, each of them employed in order to obtain different types of information about the interaction.
The **reductionist biology** is the more traditional perspective used for the PPI studies, it is focused on the analysis of the interaction between specific gene products under a molecular point of view and in a particular physiological context . It is used when is needed the analysis and the prediction of specific interactions between partners.
Protein network describes a set of interactions between proteins. It can be represented as a graph where each node (or unit) in the network is a protein, while the edges, the links between the units, represent the interactions between the proteins.
Whether the protein network is more focused on a subset of the system the **System biology** tries to look at the organism as a whole. It helps to understand the larger picture of a system (be it at the level of the organism, tissue, or cell) by putting its pieces together, it is in sharp contrast to the reductionist approach, which involves taking the pieces apart.

1.1. Principles of protein interactions and binding site properties

Any process in the cell is mediated by protein-ligand interaction, *just to have an idea of them* among them we can recognize the transduction of signals, enzyme catalysis, immune response, cellular division, programmed cell death, cell-cell recognition etc.
The protein functions are based on the capability of proteins to bind a certain ligand, where the ligand can be an elemental ion, a small organic molecule, a peptide and a macromolecule (protein, RNA and DNA).
Here a first glance of some key concepts that will be useful for understanding the PPIs.
A **binary interaction** it's an interaction and a complex between two and only two molecules, while more in general **protein complexes** describe interactions between two or more macromolecules. The **interfaces** in protein interaction are defined by all the atoms or residues that have at least one atom in contact with one atom of the partner interactor. And in general we assume that two atoms are in **contact** when the distance between them is within 5 Å.
These interaction can be **transient interactions** that are more weak interactions, they are brief and reversible, or **permanent interactions** that occur when two proteins form a stable complex. Nonetheless there are proteins that interact in a weak and transient way in one

conformation but they are able to do more strong and stable interactions when undergo a conformational change, these are the so called **strong transient interactions**.
The ability of proteins to form biologically active complexes depends on the properties of their binding surface, these are the ones that we have to take more into account:

- Size of the interfaces
- Geometric compatibility.
- Chemical composition.
- Atom packing efficiency.
- Hydrogen bond or salt bridge frequency.
- Number of buried water molecules.
- Interaction energy.
- Residue conservation.
- Types of secondary structures

The binding site (BS) properties

The forces involved in the protein-ligand interaction mostly include H-bond, hydrophobic contacts, electrostatic and Van der Waals interaction, ionic bond.
In particular geometry and electrostatic complementarity (non-covalent interaction and Van der Waals) are the two major feature for the correct match of **the binding site-ligand.**
Since the major driving forces of the electrostatic complementarity are relatively weak and act at a very short range, the two partners of the interaction have to be very close to each other and a lot of these forces are needed in order to achieve a stable interaction.
This is also accentuated by the fact that when the ligand of the interaction carries an electric charge we expect that the charge of the binding site tends to have an opposite charge, this will determine a strong **specificity** of that interaction.
The surface geometry of the protein- small ligand interactions are usually tight and deep. While when we are dealing with PPIs the surface features are usually larger and flatter**.** The area of PPIs interface is usually large (it can range from 1000 to 4000 Å^2).
Standard-sized interfaces vary from 1200 to 2000 $Å^2$.
Short-lived and low-stability complexes have a smaller interfaces (1150-1200 $Å^2$).
Protein-small molecule interaction usually have a surface area between 300-1000 $Å^2$.
Usually in PPI interfaces beta-sheets are more suitable to form protein interfaces: first of all because they can form quite a flat and extended interfaces and also because in most of the cases the composition of beta-sheet is hydrophobic.

The specific residue or the cluster of residues present on the interaction surface that make a major contribution to the binding free energy and **affinity** of the interaction are defined as **Hotspots**. These residues tend to be usually more evolutionary conserved, and they can be described as residues whose substitution by an alanine leads to a significant decrease in the free energy of binding.

Protein-ligand interaction driving force

Energy and work are necessary to keep order and stability in any interaction system. The formation of macromolecules and complexes is a process governed by the laws of thermodynamics that goes from a more disordered configuration to a more ordered one.
The thing that we are interested in is indeed the energy variation of the system during this process. To do this we will take full advantage of the so called **State function**, an object that let us measure only the initial and the final state of the system, without measuring the intermediate changes that occur during the process.
In our case the state object is the **Gibbs free energy**.
Measuring the variation of the Gibbs free energy between the final state and the initial state we can determine which is the direction of the process.

$$\Delta G = \Delta H - T \cdot \Delta S$$

In the formula of the Gibbs free energy variation the ΔH indicates the variation of enthalpy, which is the general heat released or absorbed by the system during the reaction. The variation of entropy (ΔS) represent the general tendency of a system to maximise its disorder. T is the Temperature of the system.
Important thing to remember since we are working with Biological system is that they exist in a state of **constant pressure** and **temperature** and therefore, we can neglect the temperature variation.
The variation of **Enthalpy** is a quantity that puts together the internal variation in energy (ΔE) with the pressure (P) and the volume variation (ΔV) of the system.

$$\Delta H = \Delta E + P \cdot \Delta V$$

With similar reasoning as those that we used for the Temperature in the Gibbs free energy equation, we can claim that the change in pressure and volume is constant and equal to zero because we are in a stable system for these variables. Therefore, we will have that $\Delta H \approx \Delta E$.
The **internal energy** (E) of the system is equal to the sum of the **potential energy** (U) that is defined as the energy needed to break or form interactions and bonds and the **kinetic energy** (K), the result of the molecular motions induced by the heat.
So, when a formation of a bond takes place the system releases energy and the final energy will be smaller than the initial, in this case the value of ΔE will be **negative**.
On the other hand when we break bonds we need to supply energy to the system and the resulting variation in Enthalpy will be **positive**.

The **Entropy** (S) is defined as:

$$S = k_B \cdot \ln(\omega)$$

k_B is the so called Boltzmann constant and ω is the number of possible configurations of the system (known as microstate). Entropy is generally known as an expression for the "description" of the disorder of a system. Indeed the number of possible states is inversely proportional to the order of the system.
In classic thermodynamics at constant temperature (remember that this is a feature of biological systems), entropy and enthalpy are linked from the following equation:

$$\Delta S = \Delta H/T$$

Let's now analyze what happens in **protein folding** and explain why it's a **spontaneous process**. *We will do it with the same instruments previously cyted.*
When ΔG is equal to 0 the system is at the equilibrium (and there aren't any changes).
If $\Delta G > 0$ it is an **endergonic** processes, that means that the a process requires energy in order to occur, While if $\Delta G < 0$ it is an **exergonic** processes, the system releases energy.
The general idea of the spontaneous processes is that they have to be exergonic.
Therefore, we have to analyse more in detail what happens at the ΔS (Entropy) and ΔH (Enthalpy) in the Gibbs free energy formula during the protein folding process.
The variation in enthalpy will be negative due to the formation of the electrostatic and Van der Waals interactions that tend to stabilize the system.
Then we have to consider the entropy contribute, this can have a different role depending on the analyzed process. In general the ΔS term is lower than zero because the configuration of the protein goes from a more disordered state when it is unfolded to a more ordered one when is folded. But this negative value will be summed to a minus sign into the general formula ($\Delta G = \Delta H - T \cdot \boldsymbol{\Delta S}$) leading to a positive efford.
Therefore, we have to look at the role of the **hydrophobic effect**, this element in fact guarantees the spontaneity of the protein folding process. This effect describes how the water molecule that initially are bound to the hydrophobic residues of the protein chain are released back into the solvent when these residues start to aggregate, leading to a general increase of the entropy in the system.
Summing up all these contributes we end up with a general negative ΔG that confirms the spontaneously process of the protein folding and affirm that the driving force behind many events within the cell is represented by the increase in entropy generated in the water molecules when a reaction occurs.

2. PROTEIN-PROTEIN DOCKING

In the field of molecular modelling, docking is a computational method used to predict the preferred orientation of one molecule to a second when they interact and bound to each other forming a stable complex. In protein docking we don't want to know if the two molecules interact but we want to know **how they interact**.

2.1 General principles and challenges of docking

Protein docking experiments are complex and demanding task considering all the different manner in which the two molecules can interact and all the intermolecular forces and geometrical characteristics involved.
The general idea of protein docking is indeed to generate a stable (low in energy) model through the maximization of the geometrical and electrostatic affinity between the two partners.
The docking process can be divided in two main steps:
1. **Pose generation**: it is the result of a conformational space search that tries to dock the ligand in all the possible conformations on the surface of the receptor.
2. **Scoring**: or ranking the potential solutions of all the poses of the ligand into the receptor.

In protein-protein docking jargon one of the two partner is called **receptor** and the other is called the **ligand**, this last one can be any kind of elemental ion, small molecule, peptide or macromolecules. Let's see more in detail what are the features of different class of ligands:
Protein-**small molecule**: they usually bring a lot of flexibility to the system, leading to a more demanding computational effort. The binding sites of the target tend to be small and deep.
Protein-**peptide**: generally long 8-10 amino acids. In this case there is a problem with the floppy backbones, the peptide is short and once again there is a high level of flexibility.
Protein-**protein**: the backbones are usually stable. The binding interfaces tend to be large and flat.

In order to carry out a docking experiment we would like to start with some good crystal structures, since we want to see exactly the atomic position of the interaction. Better the resolution of the structures, the more reliable the solutions will be.
If we don't have a crystal structure we are free to work with a good structure derived by the homology modelling method.
In some cases the ligand is already co-crystallised with the receptor molecule. In this case it's possible to carry out the so called **bound docking** experiment, in which starting from the existing complex once the receptor has been artificially separate from the ligand it is possible to perform the docking experiment to reconstruct the complex and basically test how reliable and efficient is the docking algorithm.
Whereas when you do the **unbound** or **predictive** modelling, you start with two separated

molecules for which you know that they interact but you don't know how they interact. Moreover, this "unbound" structure may be in its **native** form, when it's free in solution and in its uncomplexed state. It can be in the **pseudo-native** form, that means that the structure is complexed with a different molecule from the one that will be used for the docking, or **modelled**.

Another key ingredient in docking is the **representation of the system**.
The thing that you are going to model with docking is the interface of interaction, but there are different way and type for representing it:
Using the **Dense surface (Connolly)** in which the protein interfaces is represented in a very granular manner, you will end up a large and more accurate number of different solutions.
Ot the **Sparse surface (shou Lin et al.)** in which the representation of the interfaces is very rough and therefore you will end up with less accurate solutions.

Pose generation

It is a conformational space search, that can be summarized with the sentence: "getting the ligand into the pocket". The action that the algorithm have to carry out is a set of **translational** and **rotational** space movement of the ligand in respect to the receptor that instead is the stationary object.
The ranslational space movements are pretty simple, because the algorithm simply tries to put the ligand as much closer as possibile to the receptor.
What really represent an issue for the docking experiments are the torsion space movements. Considering that you can have a set of possible torsion angle for one residue with respect to the following residues, you will end up with too many valid angles and points to evaluate and this would be a huge limit in terms of time and computational power for the analysis.

There are three major types of docking algorithms for the pose generation step. In the **rigid body**, you consider both the receptor and the ligand as two rigid solid bodies. It Is faster but less accurate. The **semi-flexible docking**, in which the receptor is the rigid body (you don't allow torsions) and the ligand is flexible (torsions are allowed). This method is widely used when performing protein-small ligand docking. It is slower but more accurate. And the **flexible docking**, both molecules are flexible. But the flexibility is very limited or simplified.

The **conformational space search** needs to be carried out from an efficient search algorithm that ideally would be fast and effective in covering the relevant conformational spaces. An algorithm can be exhaustive, it basically scans the entire solutions space, and for this reason it is very demanding in time and power manner. One other solution is represented from those algorithm that use a gradual guided progression through the solution space, in which only a part of it is analysed. While the last algorithm approach is the so called data-driven docking, these algorithms incorporate available biological informations about the binding

site, putting in this way some guideline for the docking experiment it is the fastest and the most reliable one.

Ranking of potential solutions

The potential solutions of the docking procedure are then ranked using specific **scoring functions** that calculate the general energy of the interaction.
There are mainly three different classes of scoring function:
1. **Force field:** affinities are estimated by summing the strength of intermolecular van der Waals and electrostatic interactions between all atoms of the two molecules in the complex.
2. **Empirical:** based on counting the number of various types of interactions between the two binding partners.
3. **Knowledge-based:** based on statistical observation of intermolecular close contacts in large 3D databases which are used to derive "potentials of mean force".

Ideally the best search algorithm and scoring schemes should work simultaneously, with a first and fast scoring of all the possible solutions in order to obtaining in this way some initial "good" candidates, **mostly based on geometric criteria.** Followed by a more advanced methods to further discriminate the conformations that not fulfill the **energy criteria.**

Since the docking process generate a lot of different solutions, most docking algorithms return a **clustered solutions.** This means that all those solution with a similar energy in the same region of the receptor are clustered together. The cluster with the largest number of low-energy structures typically corresponds to the native complex and the center of the most populated cluster being a structure near the native binding site.

2.2 Protein-protein docking with ClusPro

In this section we will use the online tool ClusPro (https://cluspro.bu.edu/login.php) in order to perform a docking procedure of two proteins. However, it is not the only tool available for this task, there are indeed different tools that can be used and differ for the kind of docking they are most suitable for: protein-ligand, protein-nucleic acid and protein-protein.

The CAPRI (Critical Assessment of PRotein Interaction) is a community-wide initiative, inspired by CASP (Critical Assessment of Structure Prediction). The goal of fueling progress in computational methods for modeling protein complexes. It has done so by offering computational biologists the opportunity of testing their algorithms in blind prediction of experimentally determined 3D structures of protein complexes.

In the CAPRI experiment, the two web-server programs that excelled were ClusPro and HADDOCK. However, this last one is quite difficult to use, and has a very large set of

parameters that can be specified. Hence, we will focus on ClusPro.
ClusPro is the first fully automated, web-based program for the computational docking of protein structures. Its input consists of the PDB files of the 2 proteins to dock. The program firstly performs a rigid body docking, then refines it with some steps of semi-flexible docking. The output is a list of top predictions of docking conformations.

ClusPro Tutorial

In this tutorial we are going to use ClusPro to perform a docking experiment of the human MKK7 (uniprotKB AC: O14733) and Gadd45 β (uniprotKB AC: O75293) protein structures.
Gadd45 β is a key mediator for the suppression of the activity of NF-kappa B/Rel factors that control programmed cell death, crucial to oncogenesis, cancer chemoresistance and antagonism of tumor necrosis factor (TNF) alpha-induced killing. The inhibition of TNFalpha-induced JNK signaling by Gadd45 β depends on direct targeting of the JNK kinase, MKK/JNKK2. However, the mechanism by which Gadd45 β interact with MKK7 is unknown (Papa et al, JBC 2007).
Thanks to previous studies we know that the formation of the Gadd45B-MKK7 complex enables the insertion of the Gadd45 β acid loop 1 into the MKK7 catalytic pocket. This engagement of the MKK7 ATP-binding site by Gadd45 β seems to prevent access of the kinase to ATP. Therefore, the Gadd45 β-MKK7 complex formation inhibits MKK7 by preventing it from binding ATP.

Since at the time of writing the crystal structure of Gadd45B is not available in the PDB we will build the homology model of such protein using HHPred (https://toolkit.tuebingen.mpg.de/tools/hhpred) that is an online resource for homology modelling and more.
The first step is to retrieve the FASTA sequence of the Gadd45 β protein from UniProt.

>sp|O75293|GA45B_HUMAN Growth arrest and DNA damage-inducible protein GADD45 beta OS=Homo sapiens OX=9606 GN=GADD45B PE=1 SV=1
MTLEELVACDNAAQKMQTVTAAVEELLVAAQRQDRLTVGVYESAKLMNVDPDSVVLCLLA
IDEEEEDDIALQIHFTLIQSFCCDNDINIVRVSGMQRLAQLLGEPAETQGTTEARDLHCL
LVTNPHTDAWKSHGLVEVASYCEESRGNNQWVPYISLQER

This sequence will be pasted into the input textbox of HHPred to search for the suitable template of Gadd45 β. After the submission of the job HHPred will return is a list of possible templates.
A good candidate for the template can be the structure with PDB-ID: 3CG6 that has a good coverage and a high sequence identity with the target protein (Identities of 57% and Similarity of 98%) moreover a very good crystal resolution of 1.7 Angstrom.
Once satisfied with the template we will build the model by clicking on the “Model using selection” link. HHPred will then produce a model of Gadd45B in PDB format that we can

download.

To retrieve the crystal structure of MKK7 we have to go in PDB and perform an advanced search with the UniProt accession number of MKK7 (AC O14733) specifying that we want to obtain a crystal structure that has a high resolution (< 2.3 Angstrom), that is not complexed with ligand and that is not a mutant, since we want to work with the native protein. The best structure that satisfies these parameter is 6QFL.

Now we want to check if the position of the residues involved in the interaction are the same in the PDB file respect to the ones we found in the UniProt sequence and in the literature.

MKK7 residue (literature)	MKK7 residue (PDB file)	Gadd45 β residue (model)
Lys149	Lys165	Glu65
Lys157	Arg178	Glu66
Lys162	Lys173	Glu113

Table 1. Position of the interacting residues in literature/UniProt, PDB and model

Another refinement analysis before the docking procedure, is look for possible "restraints", these can enormously reduce the configuration space size and direct the search toward more likely solutions. This step can be also done directly from the "Advanced Options" menu in the web page of ClusPro and consists in the generation of a *json* file in which we will set the "required percent groups" at 75% and the "required percentage of restraints" at 100% adding then the positions and the relative chain of the interacting residues of our proteins. The distance can be set from a minimum of 1 to a maximum of 10. Then download the generated file.
Now we can go on the ClusPro page, choose a name for our Job and leave the server setting as cpu. Then upload the 6QFL.pdb file in the receptor voice specifying the chain ID and upload the Gadd45 β -model.pdb file in the ligand space also specifying the chain.
Once also uploaded the *json* file containing the restraints we can finally click on the "Dock" button to submit out job. After a while you will have your results.

ClusPro results analysis

ClusPro provides as default 10 best models for each group that respect some specific features like "model balanced", "electrostatic favoured", "hydrophobic favoured" and "VdW+Elec.". If we don't have any specific reason or previous knowledge of the interaction to choose for a specific group, the recommended set to analyse is the balanced one. Moreover the cluster 0 of each groups is usually the most populated one and the one with the lowest energy structures.

Once downloaded it we can proceed with the analysis of the reliability of the models using a visualization program like Chimera. This analysis consists on picking the distances between the residues involved in the interactions as shown in the Figure 1. If the distances are in the range of 1 to 5 Å the model can be considered meaningful in terms of interaction and docking complex generated.

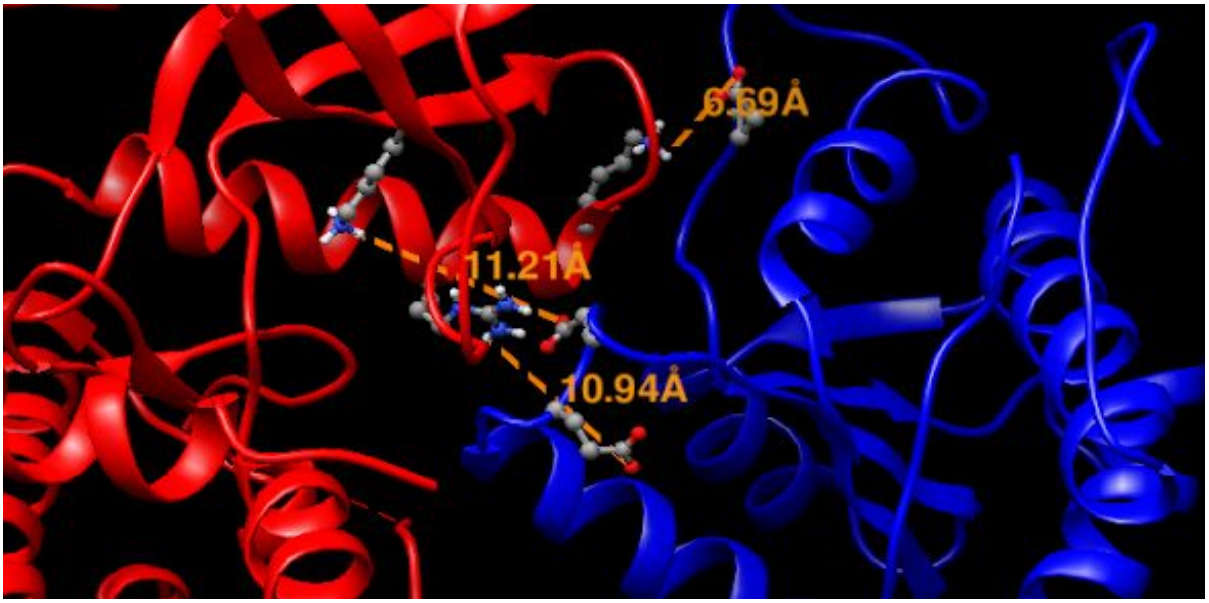

Figure 1. Docking model of Gadd45 β in blue and MKK7 in red using UCSF Chimera

As we can see from the analysis of this model none of the residue properly respect the distance for the interaction, the only one that can be close enough is the one between the residues Glu113 and Lys173 that is at 6.69 Å. While we can say that the others two pairs of residues don't interact due to their too high distances of 11.21 Å (Glu65-Lys165) and 10.94 Å (Glu66-Arg178).

3. PROTEIN-PROTEIN INTERACTION NETWORKS

Experimental and computational resources for PPI studies

Molecular interaction data can be collected from both experimental and computational approaches. Let's have a glance of the experimentally approaches commonly used to collect and analyze protein-protein interaction data.

First of all we have to make a distinction between high throughput and low throughput experimental techniques developed to generate PPI data.

High throughput techniques provide information on several thousand pairwise interaction in the same experiment and they are particularly useful and indicated to study entire interactome. On the other hand low throughput techniques produce a minor amount of data that instead are much more accurate and reliable. *They are in fact more used when a*

deeper insight into certain characteristic of an interaction is needed. Moreover they are time consuming and that means that they are more expensive.

High throughput techniques

Yeast two-hybrid (Y2H) It's the main binary method to detect direct physical interactions between proteins pairs. It is a technique based on the reconstruction of a functional transcription factor, in genetically modified yeasts strains, when two proteins or polypeptides of interest, called "bait" and "prey", interact.
Advantages: it is fats, inexpensive, scalable and it is performed in vivo system in which binding sites can be accurately mapped.
Disadvantages: there are usually a lot of false positive

Affinity purification mass spectrometry (AP-MS) It's an in vitro experiment. It is based on the separation of molecules in solution (mobile phase) based on differences in binding interaction with a ligand that is immobilized to a stationary material (solid phase). Using the Mass Spectrometer it's then possible identify the proteins of the mixture that interacted and bound with the ligand. This method allows to examine interactions among multiple proteins at very low concentration and prey proteins are present in their native state and concentration.

Medium throughput techniques

Co-immunoprecipitation (co-IP) It was traditionally considered as the gold standard assay for PPIs. A protein complex can be isolated from a protein mixture by using an antibody that is specific for one protein of the complex. In this approach the bait protein, usually expressed in the cell in vivo conditions, is affinity purified and the interacting partners are detected by mass spectrometry.

Low throughput techniques

X-ray crystallography It is considered as the gold standard for PPIs, since it provides high quality data for binding surfaces at the level of individual atoms. The protein is crystallized and the crystal is analyzed with X-ray beams. The resulting diffraction pattern are analyzed and used to reconstruct the electron density of the molecule, this undergoes through some refinement steps to reach a good quality model of the protein structure. It is an extremely challenging and expensive technique. Not every protein can be co-crystallised and some of them that c-crystallise in vitro do not interact in vivo.

3.1 Resources to study PPI

Thanks to the high throughput techniques there is a huge amount of interaction data that need to be stored and organized in dedicated databases where their retrieval and fruition is both accessible and intelligible to the user. For this very reason there is also a set of guiding principles for the data management and stewardship of databases in order to make data Findable, Accessible, Interoperable and Reusable (**FAIR data principle**).
Molecular interaction databases can be classified into three main classes, accordingly to the type of data they contain:

- **Primary databases:** contain experimentally determined protein interactions coming from either small- or large-scale published studies that have been manually curated. Example are: *IntAct, MINT, BioGRID, MatrixDB.*
- **Secondary or Meta databases:** contain experimentally determined PPIs obtained by consistent integration of several primary databases.
  Example are: *APID, PINA.*
- **Prediction databases:** combine experimentally inferred data taken from primary databases with computational predictions or molecular interactions.
  Example are: *STRING, UniHI*

Another milestone in the world of databases is the **IMEx Consortium** that is an internal collaboration between a group of major public interaction data providers who have agreed to share curation efford. At the moment it is composed by 12 active molecular interaction databases dedicated to the production of high quality, annotated date, curated to the same standards and following the same curation rules. In this way data are curated once at a single center and then exchanged between partners.

One of the major problem when trying the integration of different data stored in different databases is that each of them has its own standard, consistently different from the others. Moreover the thing is complicated from the fact that nowadays there are more or less 500 different molecular interaction databases. Dealing with this issue the PSICQUIC (Proteomics Standard Initiative Common QUery InterfaCe) initiative aims to standardize the access to molecular interaction databases, combining data residing in different sources and providing users with a unified view of these data.

IntAct database

IntAct (https://www.ebi.ac.uk/intact/) is an open-source, open data molecular interaction database populated by data either curated from literature or from direct data deposition. It is hosted by the European Bioinformatic Institute (EMBL-EBI) and it is a member of the IMEx Consortium taking also advantage from the PSICQUIC service. IntAct interface has been developed as a web-based platform in order to allow curation teams to annotate data directly into the database.

Searching in IntAct

In this brief tutorial you will learn how *to retrieve informations* on IntAct database.
The quick search panel allows to search using different types of query: protein name, gene name, Accession Number, GO term, publication ID, experimental detection method, etc.
If you search for a non-specific gene name, the database will return you a mixed set of result, but of course you can always refine your search with more accurate details by clicking on the Advanced Search button. Moreover, if you are not satisfied by the Advance Search there is the possibility to write more complex queries taking advantage of the Molecular Interaction Query Language (MIQL) that is available from the search panel.
Once you have refined your search the IntAct database will redirect you to a list of output results containing details for each interaction like the two interactors name, the "Interaction Detection Method", the source database as well as the "Interaction AC". Then you can also click on the "Customize view" to retrieve more information or directly click on the magnifying glass to have access to the complete set of details about that interaction.
The resulting interaction can obviously downloaded both in MI-TAB and XGMML format for further examinations.

STRING database

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (https://string-db.org/) is another useful online database of PPIs; the type of interactions that you can retrieve can be direct (physical) and indirect (functional) associations between proteins. This database belongs to the category of the secondary databases and the data are derived from five main sources: genomic analysis, high-throughput lab experiment, coexpression experiment, previous knowledge from public text and previous knowledge from other databases.

Searching in STRING

From the main page of STRING you can perform a search using the name (the scientific name or any accession number from other databases) or the sequence, that needs to be in FASTA format, of the protein and it's possible specify the source organism. You can also query for more protein names and sequences.
The output of the search will be a network of proteins that are involved in both primary and secondary interactions with the query. The color of the nodes and the edges depends on the characteristics of interactions as it is explained in the "Legend" view. Furthermore, by clicking on any interaction partner a pop-up will be displayed, containing information about the function, identity, structure of the protein and different links to other databases.
The plus and minus buttons allow the user to focus on smaller or larger networks changing the number of nodes. As we said it's also possible to search for a list of proteins that will limit the output network visualization to the proteins of interest and also for cluster of orthologous groups.

3.2 Principles of graph theory and protein-protein interaction networks (PPINs)

Biological networks are a useful representation of the set of interactions occurring among different proteins in a specific system and they can be described and studied with the **graph theory**.
A graph in this context is made up of **nodes** (or points) which are connected by **edges** (or links), in PPINs these elements are respectively the representations of the proteins and the interactions between them. The study of networks nowadays is very useful when dealing with a huge amount of data, and this is very appealing for biologist and bioinformaticians that want to extract information and look at a biological entity as a system.

Networks can be structured as **undirected networks**, where the nodes are connected together by edges that are bidirectional (PPINs usually belong to this category), **directed network**, where the edges point in a direction (i.e. metabolic networks, gene regulation networks) and **weighted networks**, in which the edges between the nodes bring some features, for example the measure of the reliability of the interaction or the quantitative expression change that a gene induce over another.

PPINs can be converted and mathematically described using adjacency matrices that are square matrices. Rows and columns are assigned to the nodes in the network and the presence of an edge is symbolised by a numerical value.

Let's now see some topology features of networks, these can help to identify relevant sub-structure within the network. The **degree of a node** explicate the number of edges that are incident to that point. The **shortest path** between two nodes is the path with the minimum number of edges and if the graph is weighted, it is the path with the minimum sum of edge weights. In a **scale free** network there are few nodes that have a lot of connections (these nodes are also called *hubs*) and many other nodes with low-degree of connections. The **transitivity** of a network highlights the nodes that are more internally connected than they are with the rest of the network, these topologies are also called clusters or communities. The **centrality** can be measured for nodes and for edges and it is an estimation on how important the node/edge is for the connectivity or the information flow of the network.

Protein-protein interaction networks properties

Different types of information can be represented in the shape of networks and the meaning of the nodes and edges depends on the type of data used to build the network. Based on this we can define different types of biological networks such as protein-protein interaction networks (PPINs), metabolic networks, genetic interaction networks, gene/transcriptional regulatory networks and cell signalling networks. The data used to

build these networks can be derived from manual curation of scientific literature, high-throughput dataset, computational predictions and literature text-mining. In this chapter we will focus on the protein-protein interaction networks (*PPIN*) that are a mathematical representations of the physical contacts between proteins in the cell.

One of the properties of PPINs is the so called **small world effect**. This denotes the high level of connectivity between proteins, that despite the diameter of the network in fact the maximum number of steps separating any two nodes is usually less than six. This property allows for an efficient and quick flow of signals within the network. Moreover, this feature helps the system to be extremely robust and able to deal with some perturbations. To better understand this last point we have to introduce another properties, the **scale-free** nature of PPINs. This feature describes how the majority of the nodes in these networks have only a few connections to other nodes, whereas some nodes (*hubs*) are connected to many other nodes in the network. This also reinforce the idea of the stability of the network, if for example a random failure occurs, since the majority of the proteins have a small degree of connectivity the probability that a hub would be affected is very low. Moreover, if a hub-failure occurs the network will generally not lose its connectedness due to the remaining hubs.
Another property of the PPINs is the so called **transitivity** or **clustering coefficient**, that is the measure of the tendency of the nodes to cluster together. High transitivity means that the network contains communities or groups of nodes that are densely connected internally, These communities may reflect **functional modules**, clusters of protein that even if placed in a different context preserve their intrinsic functional properties. Moreover, the identification of these modules also helps the analysis of PPINs reducing the complexity of biological networks. **Protein complexes** can be considered as a type of modules in which proteins are interacting with each other in a stable manner, maintaining a fixed configuration in time and space terms.

Building and analysing PPINs

Let's briefly explain how could be a potential workflow for building and analysing protein-protein interaction networks. First of all we have to retrieve the data, this can be done from different databases, like STRING, IntAct, MINT etc. Then we have to build the network data in dedicated programmes or tools for the visualization and analysis of the PPIN, and whether is possible identify specific structures and perform the annotation enrinìchment analysis using for example Gene Ontology (GO) or Reactome.

An important concern in network analysis is whether the interaction network can be trusted as the representation of a real biological interaction. For this reason the data of the network that we want to analyse need to be reliable and describable with a measure of confidence. The **MIscore** for example is a scoring system that allows weighing the evidence of the interaction provided by different sources and following the standards created by the IMEx

consortium.
There are many different strategies that can be used for the analysis of the topological features of a network, but in this section we will point out the **centrality analysis**. This evaluation gives an estimation on how important a node or edge is for the connectivity or the flow of informations of the network.
The **Closeness centrality** (CC) indicates how close a node is to all other nodes in the network. It is calculated as the average of the shortest path length from the node to every other node in the network and t is an estimation on how fast the flow of information would be through a given node to another node.

$$CC(i) = \frac{N-1}{\sum_j d(i,j)}$$

Where *N* is the number of nodes in the network and the denominator represent the shortest path from *i* to all other nodes *j.* the value of *CC(i)* ranges from 0 to 1.
The **Betweenness centrality** (C_B) measures how often a node occurs on all the shortest path between two nodes.

$$C_B = \sum_{j<k} g_{jk}(n_j)/g_{jk}$$

Nodes with high betweenness may have considerable influence within a network because of their control over the information passing through the network.

Another useful method for studying the PPINs is the **clustering analysis**. This is an approach that aims to reduce the network complexity for the identification of functional modules, complexes, cliques and motifs. For some of these entity like modules and complexes we've already explained the meaning. While for **clinique** we intend a subset of nodes in which every node is connected with every other member of the clinique and for **motif** a pattern of connection within the cluster that generate a characteristic dynamic response.

Studying PPI networks with Cytoscape

Cytoscape is an open source software for the visualization of molecular interaction networks and biological pathways. It also allows to integrate these networks with annotations, gene expression profiles and other state data.
For this tutorial we will use Cytoscape to create and analyse a network starting from a dataset of 15 proteins involved in parkinson disease retrieved from the SIGNOR databases.
First of all we have to import the dataset: *File → Import → Network from Public Database.*
In the text box we have to copy-paste the 15 UniProt ID of our proteins (P0CG48 P09936

P31749 Q6Y2X3 P37840 Q9BXM7 P42574 Q5S007 P55211 P98170 P42345 P99999 O60260 O43464 P0CG48). Select IntAct as the database source and then *Import*. Then manually merge the network, that in our case will be the only one from IntAct, and on the "Advanced Network Merge" menu select the UniProtKB accession numbers as a common ID for the merge.

At a first glance the network may appear crowded and difficult to interpret. Therefore, it is possible and very useful to apply a number of filters and styles to make it more readable.

For example you can filter only interaction with a confidence MIscore higher than 0.4. This can be done in the *Select* tab by clicking on the "+" and choose the *Column Filter* from the drop-down menu. Select then the voice for "Edge: Confidence-Score-intact-miscore" and the the interval between 0.4 and 1.0. You can see that the filter has been applied if some of the edges are now coloured in red. From the File menu we can create a new network using only the selected nodes: *File → New Network → From selected nodes, selected edges.*

In this new sub network the several self-loop are now visible. These may be real and important interactions, and in general it's important to retain them.

To further improve the visualization of the network we can remove the duplicate nodes:

Edit → Remove self-loops (even if the self-loop may be real and important interactions, and in general it's important to retain them). In the dialog box select Merged Network (1) and click the Ok button.

In the table you can edit the number and type of data column shown in the node, edge and network table.

Let's now build our customized filter. We want to have a network with PPIs belonging only to human. Go to the "Select" tab in the Control panel and choose "Create new filter" and give it a name, for example "parkinson-human". Then click on the "+" button and select to create a "Column Filter", and in this case we will use the node column called "Taxonomy ID". In the search bar you can type the value you want to search for. In this case we will type the "9606" that is the NCBI taxonomy identifier for the human specie. Once again we want to create a new sub network: *File → New Network → From Selected Nodes, All Edges.*

From the style panel you can customize your network interface, changing edges and nodes color, shape, label, etc. You can also modify them accordingly to some specific features in using the little black triangle.

One of the most important thing is the **Network analysis**. In order to see the topological features of the network go in *Tools → Network Analyzer → Network Analysis → Analyze Network.* In the resulting panel you can retrieve many different information on the network like the clustering coefficient, the average path length, the closeness centrality or betweenness centrality, etc.

Let's now to perform a **GO enrichment** using **BiNGO**. Before that you have to install the BiNGO app, go to *Apps → App manager* and type "BiNGO" in the search box and install it.

We will do a functional enrichment to see whether the genes of our network are related to specific biological processes.

Go to *App → BiNGO* a dialog windows will open. Choose a name of the analysis and write it

in the Cluster name box. Then paste the list of nodes of our network in the "Paste Gene from Text" box. Then: *Select ontology file → GO_Biological_process.*
Select organism/annotation → Homo sapiens and Start BiNGO.
The output of BiNGO is a table and a network. The table reports the biological processes and statistical values and the genes associated to each biological process. The network (Figure 2) represents connection between various biological processes. Each node of the graph is a biological process, colored by statistical significance (the more orange-like, the higher the statistical significance)

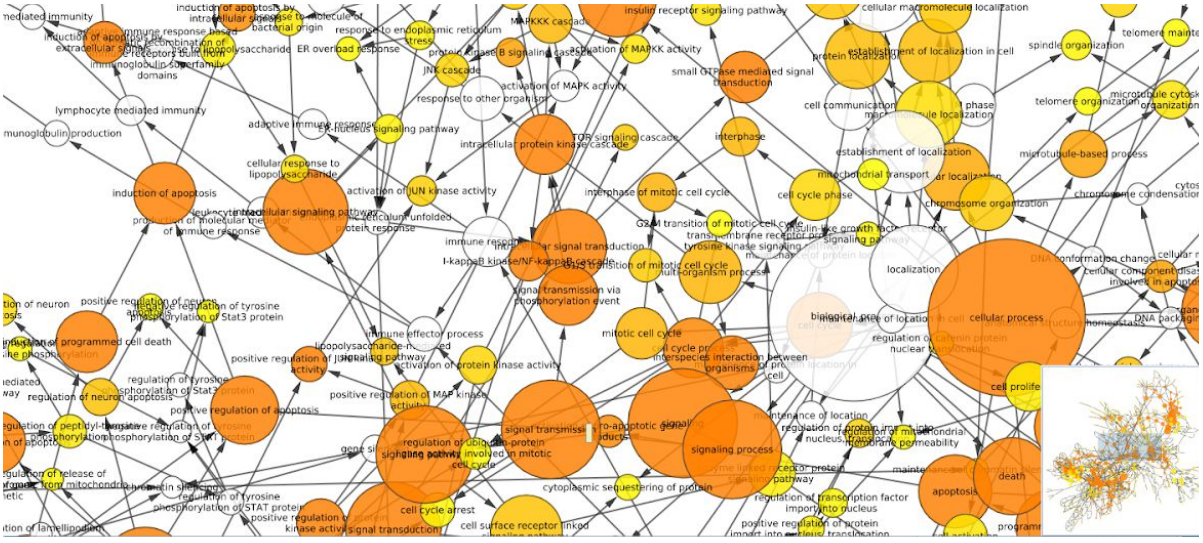

Figure 2. image of the network after the GO enrichment with biNGO in Cytoscape.

4. PROTEIN-SMALL MOLECULE INTERACTIONS

A **drug** is a substance intended for the use in the diagnosis, cure, mitigation, treatment or prevention of disease. In general a drug is a small-molecular-weight chemical compound, but can also be a biological macromolecule, such as an antibody or a recombinant protein.

4.1 Drugs and drug targets. Principles of drug design, drug target prediction and drug repositioning

While more generally a **biological target** is any molecule to which some other entity, like an endogenous ligand or a drugs is directed and/or binds, the **"druggable" target** (or drug target) is a nucleic acid or a protein known or predicted to bind with high affinity with a drug and whose **activity is modified** by the drug with a therapeutic benefit to the patient.
We will focus on proteins because they constitute more or less the 80% of all drug targets landscape both because they usually are at the surface of pathogens and host cells and because they can also act as toxins.

The **involvement of proteins in diseases** can be also represented by the those caused by the malfunction of endogenous components, for example diseases caused by the misfolding of some proteins, like cancer, metabolic disorders, etc.
The involvement of protein in disease usually can be expressed by they loss of enzymatic activity, the receptor overstimulation, the misfolding and the autoimmune response.

Drugs action can be mainly of two different types. The **Direct** action in which drugs inhibit the malfunctioning protein and the **Indirect** action in which they modulate the activity of a different protein, for compensating the abnormal activity of the malfunctioning protein.
The drugs that effect on cell-surface receptor can be **agonist** drugs that work by fully activate the receptor to which they bound and **antagonist** drugs that usually bind to a receptor but do not activate it, this action can also block the activity of other agonist. In many cases the chemical structure of agonist and antagonist is very similar and both act by "molecular mimicry", which means that they resemble the endogenous molecule acting on the surface and therefore, competing with the endogenous molecule for the binding site of the target. Another binding mechanism used by drugs is to bind the **allosteric** site of the target allowing to either activate or inhibit the enzyme activity of the target.
It is possible to design drugs that bind the target in a **reversible** way, these are usually easier to develop but at the same time since it is needed a high concentrations of the drugs for the competition that can bring to some unpleasant side effects.
As well it is possible to design drugs able to strongly bind the target. This strong interactions, that bring to an irreversible inhibition, can be exploited with the formation of **covalent bond** between the drug and the target, or without the formation of **covalent bound but with high affinity** or **suicide inhibitors**.

The very first step at the **drug development** workflow is the **drug discovery process**, that usually starts from a big set of compounds that have promising activity against a particular biological target that is important in disease. These compounds undergo through the **pre-clinical phase** that include all the trials that are not performed on humans, but in vitro or in vivo with animals. The principal steps of this process are the purification and isolation of the active ingredient, the test of the substance activity, the study of the selectivity and toxicity on cell cultures and isolated tissues and the repetition of them in animal models. At this moment only the few compound that exceeded these first steps can be tested on human in the **clinical trial.**
Since for the development of a new drug requires a huge effort in terms of time and money there is a fundamental problem with the investments in the research of drugs for rare diseases, neglected disease and new diseases. For this reason there are other approaches different from the the development of a new drug from scratch and this is the so called **drug repurposing** that involves the investigation of existing drugs for new therapeutic purpose.

The number of chemical compound that can act as drugs has been estimated to be 10^60-10^100 (chemical space). However, only a few candidate molecules enter the drug development process. So how can we identify this smaller set of molecules that will undergo the drug development process. Nowadays this is exploited with the help of computers, that filter the chemical space in order to end up with a definite and manageable group of prototypical molecules also called **lead compounds**.
There two main classes of approach among computational methods. One method predict the physiological compatibility (QSAR) while the second class predict the relative binding affinity and specificity to a target protein (**rational drug design**).
In the **QSAR** (Quantitative Structure-Activity Relationship) method the prediction consist of physico-chemical properties and biological activity of these protein. It is based on the idea that when we change a structure of a molecule then also the activity or property of the substance will be modified.

Predicting the relative binding affinity and specificity to a target protein: rational drug design

It is the process used to find new drugs based on the knowledge of a biological target. The first step in this approach is the identification of a suitable or druggable target. Then it is necessary to clone, produce and purify the target and determine the three-dimensional structure of the it. And then systematically search for small, drug-like molecules that bind to the target. When the 3D structure or the homology model of the target is known you can performed the so called **structure-based drug design** using two different method based on the type of information that are available:
1. **Ligand-based approach**: relies on the knowledge of other molecules (ligands) that bind to the biological target of interest. The group of these ligand can be used to derive the so called **pharmacophore** that is a reduced representation of the drug, including only those properties important for the desirable effect/binding on the target protein. In other words a model of the biological target may be built based on the knowledge of what binds to it.
2. **Receptor-based approach** (or **structure-guided drug design**) : relies on the knowledge of the three dimensional structure of the target and on its binding site chemical-physical features. Using the structure of the biological target, candidate drugs that are predicted to bind with high affinity and selectivity to the target may be designed.
This last approach can be exploited using the **Virtual screening** computational technique that scan the huge databases available for molecule/fragments/atoms in search of those structures which are most likely to bind into the binding pocket of a drug target. And a scoring function will be used to assess the binding affinity of each compound. The result of this procedure will be a list of compounds with different scores that fit the binding pocket of the receptor. The final step is the **optimisation** of the lead compounds, we will take those compounds that have a binding affinity in the nanomolar range and then decide which one should be used for the pre-clinical phase.

4.2 Resources and tools to study protein-small molecule interactions (tutorial/practical)

DrugBank database

The DrugBank database (https://www.drugbank.ca) is a unique bioinformatics and cheminformatics resource that combines detailed drug data with a comprehensive drug target information. DrugBank is widely used by the drug industry, medicinal chemists, pharmacists, physicians, students and the general public.
The latest release of DrugBank (version 5.1.5, released 2020-01-03) contains 13'529 drug entries including 2'630 approved small molecule drugs, 1'372 approved biologics (proteins, peptides, vaccines, and allergenics), 131 nutraceuticals and over 6'354 experimental (discovery-phase) drugs. Additionally, 5'207 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

DrugBank tutorial

From the DrugBank home page you can search for a given drug. Inside the page of that drug you can retrieve a lot of useful informations for that drug like identification, description, structure, pharmacodynamics, mechanism of action, the interaction partners etc. On the top of the page you can also look for the list of all the target of the drug action.
From *Browse* → *Pathways* you will retrieve a dynamic graph with all the interaction pathways in which your drug is involved.
Moreover from the *Search* → *Chemical Structure* you can also draw step by step the chemical structure of interest thanks to a special *drawing* tool and then search for other drugs in the database that have a structure similarity with it.
You can also search compound from their molecular weight, and perform an advanced search, in the *Help* → *Search DrugBank* there are all the indications to do this.

ChEMBL database

ChEMBL database (https://www.ebi.ac.uk/chembl) is a database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data).
The data is abstracted and curated from the primary scientific literature, and cover a significant fraction of the SAR and discovery of modern drugs. ChEMBL curators attempt to normalise the bioactivities into a uniform set of end-points and units where possible, and also to tag the links between a molecular target and a published assay with a set of varying confidence levels.

ChEMBL tutorial

From the ChEMBL database you can search for a specific term and it will return you a list of entries that match your query. These are divided by compounds, target, assay, documents, cells and tissues. The content of the entry will resume some useful information about the structure, mechanisms, metabolism and others of the compound.
In the same way to DrugBank it is available the tool for drawing the molecule and search for it based on the structural similarity. But this search, with the same parameters, will lead you to a fewer results respect to the one performed on DrugBank.