Laboratory of Bioinformatics 1 part B - Allegra Via

Saul Pierotti

March 20, 2020

Protein interaction and properties of binding sites

- The binding sites for small molecules tend to be small and deep
- Protein-protein interactions (PPIs) can be approached in from different perspectives
  - The reductionist approach focuses on specific molecules of interest and analyses specific interactions
    - It aimes at predicting parteners and modes of interaction
    - It requires wet-lab experiments
  - The protein network approach focuses on a set of proteins that interact with each other
    - It relies on the underlying biology and focuses on predicting new interactions
    - It can identify hubs of proteins that frequently interact with each other
    - A singleton is a protein that has only one interaction partner
  - The system's biology approach uses matematical models that rely on differential equations
    - The inputs to the equations are parameters that describe the biological system, such as pH and redox potential
    - It focuses on perturbation to the system
    - It can rely on experiment to determine the effect of a perturbation
- Interactions can be among proteins, with other non-protein macromolecules or with small molecules
  - In this course we will focus on interactions among proteins (PPIs)
- Several ways to classify PPIs
  - Interactions can be direct or indirect
  - Binary interactions or protein complexes (2 or many interactors)
- The protein binding interface is defined as the set of atoms of a protein which are less than 5 Å apart from an atom of the partner protein
- PPIs are an essential part of signal transduction, enzimatic activity, cellular division, metabolic networks
- Protein ligands can take part in catalitic mechanisms, regulation, and other biological activities
- PPIs can be classified depending on the biological context, binding permanence, similarity of binding partners and number of partners
- Transient interaction persist for a short time, while permanent interactions are more persistent
  - Transient interactions are typically mediated by short linear motives, PTMs, disorder-to-order transitions
  - Stable interactions form homo- and hetero-oligomeric protein complexes
- Conformational changes can alter PPIs (!)
- The interactome is the set of interaction in a give compartment

Driving force of protein-ligand interctions

- To maintain the order observed in biological systems, work is required
  - Protein folding requires work to be accomplished, since it requires a local entropy reduction
  - The synthesis of macromolecules requires the energy deriving from ATP hydrolsys
- The spontaneity of chemical reactions is expressed in terms of variation in Gibbs free energy (ΔG)
  - It is a state function: it does depend only on the initial and final state of the transition, not on

the path followed
  - It represent the energy of a system that can be used to produce work
  - It can be related to 3 other state functions: temperature (T), entropy (S) and entalpy (H)
    * $G = H - TS$
  - A decrease of G (a negative ΔG) indicates a spontaneous process
- Biological systems exist in a constant pressure and constant temperature environment: we can neglect changes in temperature
  - $\Delta G = \Delta H - T\Delta S$
- The entalpy of a system is a term that condenses its internal energy (E), volume (V) and pressure (P)
  - $H = E + PV$
- The internal energy E is composed of a kinetic and a potential term
  - $E = U + K$
- Biological systems are typically in a solid or liquid state: we can ignore volume and pressure changes
  - $\Delta H \approx \Delta E$
  - This is valid for bond formation/breaking, variation in weak interactions, variation in atomic motion indiced by heat
- Since we are at constant pressure, variations in entalpy equate heat transfer with the environment (Q)
  - $Q_p \approx \Delta E \approx \Delta H$
- When bonds are broken, energy is released: $\Delta H > 0$
- When bonds are formed, energy is absorbed from the environment: $\Delta H < 0$
- Entropy (S) is a quantity related to the number of microstates (Ω) that correspond to a given macrostate
  - $S = K_b \ln \Omega$
  - A microstate is a unique atomic configuration
  - A macrostate is an observable state of the system that can correspond to many unique atomic configurations
  - At constant temperature we observe that $\Delta S = \frac{\Delta H}{T} = \frac{Q_p}{T}$
- A process is at equilibrium when $\Delta G = 0$, endoergonic when $\Delta G > 0$ and exoergonic when $\Delta G < 0$
- Protein folding is spontaneous since the entropy reduction of the protein is more than offset from the entropy increase of the solvent
  - Entropy increase of water molecule of the solvent is a driving force in many biological processes
- Ligand binding is driven by the hydrophobic effect and Van der Waals interactions
  - PPI interfaces tend to be hydrophobic
  - The binding of small molecules is typically guided by geometry of the binding pocket and elecrtrostatic interaction
- The size of a PPI surface is related to the binding strenght
  - Standard interfaces are 1200 to 2000 Å^2
  - Low stability complexes have interfaces of 1150-1200 Å^2
  - Small molecules interact with proteins in a 300-100 Å^2 area, typically in deep pockets
- PPI interfaces tend to be flat and without pockets
  - The center of the interface tends to be particularly conserved
- Transient interfaces tend to be smaller, more hydrophobic and with better complementarity
- Homomeric interfaces resemble protein cores while heteromeric interfaces look more like non-binding surfaces
- The specificity of binding is given by electrostatic interactions
  - These also prevent aggregation
  - Specificity is low if a protein can bind many partners
  - PPI affinity ranges from pM to mM, but they are typically quite specific
- Some binding energies to be rembered
  - General range: -2.5 to -22 kcal/mol
  - Interactions in signal transduction are weak
  - Cofactor binding: -5.5 gto -9.5 kcal/mol
  - Antigen-antibody: -5 to -11 kcal/mol
  - Enzyme-inhibitor: -9 to -15 kcal/mol
  - Enzyme-transition state: -17 to -27 kcal/mol

- Interaction hotspots are PPI interfaces that are quite heterogeneous
  - The binding strenght is typically given by a few hydrophobic residues (hotspots)
  - They can be identified by alanine replacement: if we replace an hospot with alanine the affinity changes of at least 2 kcal/mol
  - Alanine replacement can also be done in silico (alanine scanning)
- Hotspots tend to account for less than 50% of the interaction surface, are quite conserved, appear in clusters
  - They are frequently represented by aromatic residues (F, Y, W)
  - They are surrounded by less important residues that shields them from the solvent

Protein docking

- It is biologically important to understand the molecular position of a ligand on a protein (its pose)
- Docking finds the optimal interaction, but cannot determine if the interaction actually happens or if it has a biological meaning
- Molecular docking is an optimization problem: we want to maximize the interaction of the ligand with the protein by operating on their torsion angles
  - We want to maximise electrostatic and geometric affinity
- We first perform a step called pose generation that explores the conformational space, and subsequently we rank the possible solutions
- A ligand can be a small molecule or also a macromolecule
  - Binding sites for ligands are small and deep
  - Peptides are 8-10 aminoacid long and tend to have a floppy backbone
  - Proteins tend to have rigid backbones: interfaces are large and flat
- Docking is a golden standard in PPIs, because it is reliable, but it is computationally expensive: it is not an high throughput method
- A must for docking is to have a good 3d structure or model
  - Usually a crystal structure is better than an homology model, but also a crystal is a kind of model, it can be wrong (!)
- In order to test a docking method, I can take a structure with a co-crystalised ligand
  - I artificially separate the molecules and compare the docking prediction with the experimental data
  - This method is called bound docking
- On the contrary, in unbound docking it may be that I am using a structure for a protein that is in the wrong conformation for the interaction
  - It is far more complex than bound docking
  - A structure is defined native if in an uncomplexed state, pseudonative if complexed with a ligand different from the one used in the docking
  - I can also do unbound docking on a model
- Typical limitations of the apprach are: conformational changes, errors in the structures or models, limited computing resources
- In order to do docking I need a representation of the protein surface
  - The kind of representation that I use will influence the algorithms that I can employ for the docking
  - I can use mathematical models that describe its shape and electrostatic proprieties
  - The protein frame can be treated in a static or dynamic way
- I can consider the protein and its ligand as rigid or flexible
  - Rigid docking can explore the whole protein surface and it is faster, but less accurate
    - Rigid docking algorithms proved to be useful for enzyme-inhibitor and antigen-antibody docking
  - Semi-flexible docking treats the ligand as flexible and the protein as rigid
  - Flexible docking treats everything as flexible: it is more accurate but really slow
- The search algorithms can work in different ways, but it produces an enormous number of solutions

(10^9)
  - Brute-force the solution space
  - Guided progression through the solution space, sometimes only considering solution that confrom to pre-determined criteria
  - Data-driven: uses the available information about interface residues
- Scoring functions are needed for ranking the possible solutions, and they can use different approaches
  - They can use a force field, a set of parameters that define the potential energy of a system and it is based on molecular mechanics
  - Empirical functions are simpler and use a reduced description of the most important interactions
- The initial screening of poses is usually based on geometric criteria, then top-ranking conformations are discriminated with more advanced methods based on energy contributions
- The native pose tends to belong to a cluster of high-scoring poses
  - Events that occur in cluster tend to not be random
  - The native pose is typically at the center of the most populated cluster of solutions
- CAPRI is an international competition for docking algorithms organised by EBI, like the CASP is for structure prediction
- ClusPro and HADDOCK are 2 docking web-servers

PPI data and databases

- There are databases that collect only experimental data, and databases with computationally-derived interactions
- Experimental data are not always more reliable than computationally derived ones (!)
  - There are no experimental method that can replicate binary interactions under physiological conditions
- Low-thoroughput techniques can be more reliable of high-thorughput ones
- High-troughput techniques: yeast-2-hybrid, affinity purification MS
- Medium-troughput techniques:
- Low-troughput techniques:
- Methods can be binary or co-complex
  - Co-complex methods can detect non-direct interactions among proteins, i.e. if they belong to the same complex
- Yeast-2-hybrid (Y2H) is fast and scalable, but has disadvantages
  - An yeast protein can act as a bridge and give a false interaction
  - The conditions in yeast can not be physiological
  - The proteins that I am studying could in reality reside in different compartments
- Affinity purification MS: affinity chromatografy and the MS
  - It is in vitro (!)
  - I cannot identify proteins in really low amounts by MS, and proteins that are not in databases
  - False positives due to breakage of cell compartments
- Co-immunoprecipitation: antibodies interact with protein A, that interacts with protein B
  - With the antibody I can pull down protein B
  - Similar pros and cons with affinity purification MS
  - The interactors can be studied by MS
- X-ray crystallography: high detail level but challenging
  - Very low troughput
  - Artificial environment
  - There are false positives and negatives
- Data must be accessible and intellegible to the user
  - It requires a lot of effort to implement, curate and maintain PPI databases
- In the past data producers where also data analyzers
- Today this is not possible any more: specialized jobs and not enough time and resources for data producers

- Data should comply with the FAIR principles
  - Findable, Accessible, Interoperable, Reusable
  - We want unique and persistent data identifiers
- Primary databases: experimental interactions manually curated
  - IntAct, MINT
- Secondary (meta) databases: integration of primary databases
  - APID, PINA
- Prediction databases: integration of experimental predictions with computational predictions
  - STRING
- IMEx is an international consortium of interaction data providers who share curation efforts
  - Data is curated once and then shared among members
  - Curation is entirely manual
  - Data is provided in standard formats: MITAB or PSI-MI XML 2.5
  - UniProt, MINT, IntAct are in the consortium
  - STRING is NOT in IMEx because it includes predictions
- The ideal would be to have a single database to uniform interface and data mining, but this is not possible
- A good compromise: standards and guidelines to uniform data access and retrieval
- The problem of standards: creating a standard to uniform previous competing standards create just one more competing standard
- PSIQUIC is an unified query client interface for retrieving data from some biological databases
  - It provides links to the included databases
- Mentha and virus-mentha are interactome databases

Interaction networks

- Material was extracted from the EBI online course
- Networks are everywhere
- They where first studied in the context of social networks
- Euler first represented a problem with a graph: Koenisberg 7 bridges
- Networks are useful when dealing with large amounts of data
- Network biology is a branch of system biology
- Indirected networks: no direction in the interactions, es PPI networks
- Directed networks: unidirectional interaction, metabolic networks and gene regulation
- Weighted networks: edge encode a weight such as the reliability of an interaction, the quantitative expression induction, sequence similarity of genes
- A graph can be represented by an adiacency matrix
  - 1 means connected, 0 not connected
  - In an undirected network the matrix is symmetric, in a directed network it is not
- The topology of a network refers to its connectivity
  - A topological cluster is more connected to itself than to other parts of the network
  - The degree of a node is the number of edges connected to it
  - A network is scale-free if it has a small number of high degree nodes and many nodes with low degree
  - The degree centrality of a node tells us how relevant is a node for the entire network
- Our knowledge of interaction networks is noisy and incomplete
- Some features about a network could change in the future, when more data are available
- Small world effect: the shortest path among any 2 nodes tends to be small: 6 degrees of separation
  - Signals are quick and efficient, and reliable
  - Perturbating a single node does not cause enormous alterations on the network
- PPINs are tendentially scale-free: many nodes with few connections and a few hubs
  - A random mutation in the network is unlikely to hit an hub
  - It is vulnerable to targeted attack on hubs

- PPINs have high transitivity: many clusters with many internal interactions
- The PSI-MI score assigns the reliability of interactions
  - It evaluates the number of independent publications advocating for an interaction, the type of experiment that supports it and the interaction type
  - Interaction type refers to direct or indirect
  - It is a normalized score so it ranges from 0 to 1, above 0.4 the interaction is quite reliable
- The study of centrality of a network tries to understand which nodes/edges are essential for the functioning of the network
  - Drug targets are typically centralities of their networks
- Centrality can be measured in different ways
  - Degree of the node: it is a local measure that ignores the structure of the network
  - Global centrality and betweennes centrality: global measures
  - Random walks: average distance of the path from a random node to the one of interest
- Closeness centrality: how short is the shortest path from node i to any other node
  - $CC(i) = \frac{N-1}{\sum_j d(i,j)}$