Structural prediction of protein-protein interactions by docking: application to biomedical problems

Didier Barradas-Bautista[1], Mireia Rosell[1], Chiara Pallara[1], Juan Fernández-Recio[1,2*]

[1] Barcelona Supercomputing Center (BSC), Joint BSC-CRG-IRB Programme in Computational Biology, Barcelona, Spain

[2] Structural Biology Unit, Institut de Biologia Molecular de Barcelona, CSIC, Barcelona, Spain

*Corresponding author: e-mail address: juanf@bsc.es

Barcelona Supercomputing Center

Jordi Girona 29, 08034 Barcelona (Spain)

Funding: Spanish Ministry of Economy grant number BIO2016-79960-R; DBB is supported by a predoctoral fellowship from CONACyT; MR is supported by an FPI fellowship from the Severo Ochoa program.

ABSTRACT

A huge amount of genetic information is available thanks to the recent advances in sequencing technologies and the larger computational capabilities, but the interpretation of such genetic data at phenotypic level remains elusive. One of the reasons is that proteins are not acting alone, but are specifically interacting with other proteins and biomolecules, forming intricate interaction networks that are essential for the majority of cell processes and pathological conditions. Thus, characterizing such interaction networks is an important step in understanding how information flows from gene to phenotype. Indeed, structural characterization of protein-protein interactions at atomic resolution has many applications in biomedicine, from diagnosis and vaccine design, to drug discovery. However, despite the advances of experimental structural determination, the number of interactions for which there is available structural data is still very small. In this context, a complementary approach is computational modeling of protein interactions by docking, which is usually composed of two major phases: i) sampling of the possible binding modes between the interacting molecules, and ii) scoring for the identification of the correct orientations. In addition, prediction of interface and hot-spot residues is very useful in order to guide and interpret mutagenesis experiments, as well as to understand functional and mechanistic aspects of the interaction. Computational docking is already being applied to specific biomedical problems within the context of personalized medicine, for instance, helping to interpret pathological mutations involved in protein-protein interactions, or providing modeled structural data for drug discovery targeting protein-protein interactions.

Keywords: Protein-protein interactions, complex structure, computational docking, interface prediction, hot-spot residues, drug discovery, edgetic effect, pathological mutations.

1. Importance of protein-protein interactions in cell

A cell is the basic structural and functional unit of any living organism. From single cell organisms to multicellular organisms, most of the cells have information stored in the DNA, coded in the form of nucleotide sequences, which must be transcribed into RNA, and then in turn into a chain of amino acids, the building blocks of proteins. This straightforward flux of information is the so-called “central dogma” Crick (1970). However, this linear view of the flow of information is incomplete. In nature, self-interacting elements capable of modifying the above described flux of information challenge the idea of the central dogma. This is the case of ribozymes with self-catalytic activity (Lilley & Eckstein, 2007), and prions (Derkatch & Liebman, 2007), misfolded proteins that can alter the structure and function of other proteins. These self-interacting elements add loops to the straight line in the central dogma. Even with these added loops, this view does not fully depict the crowded and dynamic environment inside the cell. There are additional genetic mechanisms that regulate the levels of proteins. An example of this is the field of epigenetics where the marks found in the DNA nucleosomes, such as methylation, prevents the transcription of DNA (Bharathy & Taneja, 2012). Proteins themselves appear to have an active role to protect the balance of gene products when the cell presents an abnormal load of the genetic material like in polyploidy Stingele et al. (2012). Among all of the interactions and factors that are driving all these processes, proteins have a prominent role as they can serve as scaffolds, provide protection to RNA or DNA (chaperones and nucleosomes), and act as receptors or effectors (such neuropeptides and enzymes).

Most proteins do not act as isolated units, and their interactions with biomolecules, including other proteins, are essential in the virtual totality of cellular events (Stingele et al., 2012; Teichmann, 2002). The majority of cell processes require the assembly of protein complexes, which constitute the so-called quaternary structure.

The relationship between the genetic information contained in the DNA and the structure of proteins is currently object of intense investigation. Recent sequencing efforts have yielded much information on the variants in genes (mutations), and association studies have revealed that these variations are tightly linked to the physiological outcome of the organism (Freedman et al., 2011; Lander, 2011). There are two major approaches to analyze the effect of these variants: a reductionist view where the analysis is focused on the molecular effect of a mutation based on the 3D atomic structure of the protein of interest, and a systems approach focused on the effect on the network generated by the interactions between the elements in the cell (Figure 1). The synergy between these two approaches provides understanding on how variations in the genetic information can have effects on the phenotype ranging from an atomic level to the entire network organization, and for this, understanding protein-protein interactions from structural, dynamics and energetics points of view is essential.

[INSERT FIGURE 1 HERE]

2. Protein-protein interactions and human disease

2.1. From gene to disease: Towards personalized medicine

High-throughput techniques, like genome sequencing, mass spectroscopy and DNA and RNA expression microarrays, are dramatically changing the way we study biological sciences.

The first major change arises from the massive data generated by these techniques. Next-generation sequencing (NGS) technologies have dramatically lowered the costs of gene sequencing, and are providing genomic information for an increasing number of healthy individuals and patient populations. A biological scientist has to face the overwhelming stream of information from different sources, ranging from microorganisms (Venter et al., 2004) to patients in health care systems (Baoying, Ruowang, & William, 2015). Computational resources are fundamental to efficiently analyze all this data. Institutes like National Center for Biotechnology (NCBI) and the European Bioinformatics Institute (EBI) receive data from different sources and store it in big public databases such the GenBank (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2005) and UniProt (UniProt, 2007). Moreover, they have integrated a variety of tools like BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) or CLUSTAL (Higgins & Sharp, 1988) in publicly available websites with the goal of providing the scientific community with analytical tools for their research. This vast amount of information is an opportunity for biological sciences to put statistical methods and rigorous mathematical models into the molecular details that rule a living organism.

Following the first human genome completion (Human Genome Sequencing, 2004), the scientific community started an international effort known as the “1000 genomes project” (Genomes Project et al., 2015). The project, now finished, consisted in obtaining the genome sequence from subpopulation around the world, making the genomes available to the scientific community for a variety of analysis. It also provides a framework for important questions on human genetics. In the past, the study of the genetic variation in the human population or genotypes was only possible using the unique gene variants that gave rise to evident distinct states or phenotypes. While the term genotype refers to the information stored in the DNA

sequences, the term phenotype refers to the product of the genotype or “what we can see”, which can mean a protein fold or a cell type or even the look of the whole organism. With the lower costs of genome sequences and resources like the "1000 genomes", common genetic traits were found to be present in a large proportion of the human population (International HapMap et al., 2010). Many of these traits were determined by Single Nucleotide Polymorphisms (SNPs). SNPs are single base pair changes in the DNA sequence that occur with high frequency on the human genome (Genomes Project et al., 2010) and the field of human genetics now use it as the unit for genetic variation in populations. The International HapMap Project aims to identify changes among the genomes and to find correlations with the observed phenotypes. The number of SNPs per human genome is estimated to be around 10 million, all of them showing a different effect. HapMap has so far catalogued 1.6 million SNPs with genotypes from 11 human populations, including Japanese population from Tokyo, the Yoruba population from Africa, Han Chinese from Beijing, and European descent population (International HapMap et al., 2010; Ritchie et al., 2010; The International HapMap, 2005).

Genome-Wide Association Studies (GWAS) are a powerful tool to identify a link of a relevant SNP with a human disease (Welter et al., 2014). The goal of GWAS is to identify genetic risk factors through various association tests, backed by statistical analysis, to make predictions about who is predisposed to a given disease, and then determine the genetic interplay of disease susceptibility for the development of new therapeutic strategies (Bush & Moore, 2012). The most successful application of GWAS has been the identification of DNA sequences that play a role in drug response (metabolism, efficacy or adverse effect). Warfarin dosage is an obvious example of this success (Cooper et al., 2008). A GWAS study led to discover a set of SNPs in

several genes that influence warfarin dosing. This, with further validation studies, became a clinical genetic test, which allowed physicians to give the correct amount of warfarin to patients.

The relationship between genetic analysis and clinical outcome fostered the field of personalized medicine. The current project "10K genomes" in the United Kingdom (Koepfli, Paten, Genome, & O'Brien, 2015) is a scientific enterprise taken by the British government for a personalized medicine in the public health care. The objective is to diagnose patients with rare diseases, who otherwise would never get proper treatment. Candidate genes detected through GWAS are generating large datasets of genetic variants associated with disorders, which are being deposited in public databases, such as Online Mendelian Inheritance in Man (OMIM) (Scott, Amberger, Brylawski, & McKusick, 1999), the database of Genotypes and Phenotypes (dbGAP) (Tryka et al., 2014) or Humsavar (UniProt, 2007).

2.2. The human protein-protein interactome: A link between gene and system

The analysis of the data obtained by high-throughput technologies also produced a revolution in the biological field. It marked the start of the “OMIC era” (Kandpal, Saviola, & Felton, 2009). Genome, Proteome, Peptidome, Exome, Transcriptome, are different ways to profile and classify the biological activities of the cell. However, the analysis of any of these profiles in isolation does not give the answer to fundamental questions about the genotype-phenotype relationship (Vidal, Cusick, & Barabasi, 2011). To infer the physiological effect caused by the changes in these profiles is necessary to study how the elements of a cell affect each other. Such “omic” sciences require an integrated approach to study the elements on a given condition by analyzing the interplay between these elements to achieve a biochemical function within the context of a network (Wu, Hasan, & Chen, 2014). The signaling pathways of the cell constitute a well-

understood example of how the elements of the cell interact to elicit a molecular process. From an outside stimulus, receptor proteins transduce the signal using small molecules known as second messengers, such as the circular Adenosine Monophosphate (cAMP). Enzymes like kinases use the energy stored in Adenosine Triphosphate (ATP) to activate other proteins and start a cascade that produces the release of other second messengers, like Inositol Triphosphate (IP3) and calcium ions. Second messengers can be sensed by other proteins to inhibit the signaling or to start other pathways, in many cases reaching the nucleus and regulating the DNA transcription (Lemmon & Schlessinger, 2010). Pathways become interconnected networks when components of one pathway interact and control elements of another pathway. Graph theory can help to analyze a system as complex as the cell. A young discipline in biology, Systems biology, is taking advantage of computational approaches to understand how these interactions can have a response (Ma'ayan, 2009). Systems biology is the study of how molecules interact to give rise to subcellular machineries that form the functional units capable of performing the physiological functions needed for the cell, tissue or organ (Bhalla & Iyengar, 1999). The network analysis in systems biology intent to gain biological meaning using a global network diagram derived from available data (Wu, Harrison, & Chen, 2009).

Large-scale studies at proteomic level have become widely accessible to the community (Chuang, Kozakov, Brenke, Comeau, & Vajda, 2008; Kuhner et al., 2009; MacBeath, 2002) and are generating a diverse and increasing amount of data, including protein binding and pathway information (Aranda et al., 2010; Ogata et al., 1999; Szklarczyk et al., 2015). This has facilitated the computational construction of genome-wide networks of interactions, or "interactomes" (Rolland et al., 2014). Thus, a system-wide approach can point out the essential elements for regulating a given biological process (Wu et al., 2014). For example, the response to a stimulus

depends on the state of the signaling networks, and this can be used in system biology to predict the outcome of such stimulus at molecular level (Janes et al., 2005). An interactome network describes the interaction of genes or gene products, which means that to provide some explanation of the genotype-phenotype relationships the networks have to include interactions at different levels. To make the predictions reliable and unbiased, the macromolecular interactions such as DNA-protein, post-translational modification and its target, or protein-protein interactions (PPI) need to be of high quality and extensive (Rolland et al., 2014). PPIs are probably the most critical networks as they underlie in almost all key cellular events like proliferation, cell signalling, regulation or cell morphology alteration (Teichmann, 2002).

The most widely-used high-throughput laboratory techniques to construct PPI networks are perhaps the Yeast Two-hybrid (Y2H), and Tandem Affinity Purification coupled with Mass Spectrometry (TAP-MS). Y2H is an ingenious system that uses separable transcriptional factors and a reporter gene to prove the interaction between two proteins. The transcriptional factors have two separable domains, a DNA-binding domain (BD) and a transcription activation domain (AD). The target protein is fused with the BD and is called the bait, the binding partner is fused with the AD and is called the prey. The interaction between bait and prey reconstitute the function as a transcription factor, which can allow the expression of reporter gene downstream from the AD binding sequence (Fields & Song, 1989). TAP-MS relies on tags attached to the N-terminus of target proteins. The intended target proteins are expressed inside the cell and allowed to interact. Then, the target protein complexes are isolated by two steps of affinity purification. The proteins that co-purified with the tagged proteins are identified by mass spectrometry (Puig et al., 2001). Complementing the initially constructed networks with text mining of the literature has facilitated building the interactomes of different organisms, like S. cerevisiae (Ito et al.,

2000; Uetz & Hughes, 2000), C. elegans (S. Li et al., 2004), A. thaliana (Cui et al., 2008), D. melanogaster (Giot et al., 2003; Guruharsha et al., 2011) and human (Ewing et al., 2007).

The estimated size of the human interactome ranges from 130,000 to around 650,000 binary protein-protein interactions (Rual et al., 2005; Stumpf et al., 2008). Among them, the number of protein-protein interactions that are known with high confidence ranges between 14,000 PPIs (Rolland et al., 2014) and 93,000 PPIs (Interactome3D January 2017 release; http://interactome3d.irbbarcelona.org/), which shows that the human interactome is far from being completed. The main challenge in the study of the interaction networks is to extract biologically relevant information from an extensive list of interactions taking into account different sources of the data, in order to gain insight into the molecular mechanism that drives various conditions (Glazko & Emmert-Streib, 2009; Khatri, Sirota, & Butte, 2012). Comprehensive integrative approaches that take into account data from DNA microarrays, protein expression, PPI information, and interaction with metabolites are added to the complexity in the analysis of cellular functions (Ideker et al., 2001; MacBeath, 2002). To gain knowledge from this vast source of information, network and pathway analysis can help to interpret the changes in the PPIs caused by external stimuli. The first generation of human protein interaction sets allowed network-based answers to the genotype-phenotype relationship, however, given their limited quality were not useful to make global, accurate interpretations (Rual et al., 2005; Stelzl et al., 2005). Network analysis used the topology of the network to highlight key nodes and strong interactions between different molecules, known as modules (Hartwell, Hopfield, Leibler, & Murray, 1999; G. Li et al., 2014). In network analysis, biological networks are described as “small world and scale-free” (Barabasi & Oltvai, 2004). This basically means that the human interactome contains several highly connected molecules, i.e. nodes that are known as

“hubs.” These proteins usually have a fundamental role in signaling pathways and their function is almost essential for the cell. The highly dynamic character of the interactions in the signaling pathways is a characteristic that provides robustness to the interactome (Albert, Jeong, & Barabasi, 2000).

In complex networks like the human interactome, there are no clear clusters because of the scale-free property. The scale-free property makes biological networks similar to nonlinear problems like chaos, phase transitions, and fractals (Strogatz, 2001). In fact, using only topological information and a nonlinear dynamical modelling known as the ant colony optimization, revealed fractal-like patterns in protein interaction networks in yeast (Wu & Chen, 2012), Breast Cancer (Wu, Harrison, et al., 2009), and Alzheimer disease (Wu, Huan, Pandey, Zhou, & Chen, 2009).

This indicates that the complexity of the PPI networks changes in a continue manner due to the dynamics of the cell. On the other hand, we know that activity in a cell emerges from functional modules, defined as a group of different proteins that interact but that are not necessarily present in the same space and time (Hartwell et al., 1999; Pizzuti & Rombo, 2014). Thus, there must exist some degree of clustering. There are two different ways to detect functional modules: graph clustering, or distant-based clustering. Graph clustering takes full advantage of the topology itself, as it searches for groups of nodes in the network that have more intra-connections than inter-connections. Some graph clustering methods are Highly Connected Subgraph (HCS) (Hartuv & Shamir, 2000), Restricted Neighborhood Search Clustering (RNSC) (King, Przulj, & Jurisica, 2004) and Markov Clustering (MCL) (Enright, Van Dongen, & Ouzounis, 2002). In the distance-based clustering method, some metrics from graph theory become the similarity measure that clustering algorithms will use to identify the modules. Some

of these metrics are the number of edges (Vazquez, Flammini, Maritan, & Vespignani, 2003), shortest path (Arnau, Mars, & Marin, 2005), and shortest path profiles (Maciag et al., 2006).

2.3. Interaction networks are key to understand biological pathways

Parallel to the network analysis, pathway analysis is a simplified approach that reduces the complexity of interpreting all available data and increases the explanatory power. Grouping proteins, genes, and PPIs according to the biological process where they participate can reveal clustering for a given event. This categorization breaks down long lists into smaller subsets that can be used to identify differences between two conditions, thus increasing the explanatory power (Glazko & Emmert-Streib, 2009; Khatri et al., 2012). Pathway analysis is different from the network analysis, because it uses functional information about the proteins, like cellular localization, catalytic activity, and processing aspects. Pathway analysis is more successful when it includes PPIs networks, Gene Ontology terms (GO) and expression data. The assumption that proteins in the same pathway and with common functions are tightly regulated can lead to the discovery of the “pathway network module”. In this way, we can delimit a large set of proteins that co-regulate each other to perform a particular cellular function (Wu et al., 2014). Additionally, in some biological networks, there is a correlation between GO terms and node distance (Y. R. Cho, Hwang, Ramanathan, & Zhang, 2007; Lord, Stevens, Brass, & Goble, 2003; Sevilla et al., 2005). On the downside, the annotation of a GO term has a heterogeneous origin, based on a variety of experiments and computational methods, which often leads to inaccurate/contradictory annotations and interpretation problems due the functional diversity of the proteins under different conditions (Luciani & Bazzoni, 2012).

There are different databases for protein networks and biological pathways: Biogrid (Chatr-Aryamontri et al., 2017), Reactome (Croft et al., 2011), KEGG (Qiu, 2013), STRING (von Mering et al., 2003), PAGED (H. Huang et al., 2012), HPD (Chowbina et al., 2009), BioCarta (Nishimura, 2001), or Interactome3D (Mosca, Ceol, & Aloy, 2013). Many of these databases provide, in addition to the list of interactions, information like the effect of the interaction (inhibition or activation), or the location of the interaction (e.g., nucleus, cytoplasm, and so forth). On the other hand, a number of databases provide experimentally obtained structures of PPIs but lack the integrating context of the networks: 3D interologs (Lo, Chen, & Yang, 2010), 3D complex (Levy, Pereira-Leal, Chothia, & Teichmann, 2006), SCOPPI (Winter, Henschel, Kim, & Schroeder, 2006), IBIS (Shoemaker et al., 2012), 3did (Mosca, Ceol, Stein, Olivella, & Aloy, 2014), PIFACE (Cukuroglu, Gursoy, Nussinov, & Keskin, 2014). Interestingly, STRING and Interactome3D provide the 3D structures of the proteins and the complexes they form, in the context of network data.

2.4. Disease-related interaction networks

Smaller subsets of the human interactome can be used to find answers to the genotype-phenotype relationship. Combining GWAS data, technically a “cause-effect” list for genes, with the network view has provided the most comprehensive data for complex diseases. As complex diseases are caused by several genes (e.g., heart disease, cancer, and diabetes), the use of networks seems a natural approach to gain insight on their molecular basis. The human diseasome, which links phenotypic features to all known disease genes, is the result of that approach (Goh et al., 2007). The human diseasome can be exemplified by a bipartite graph in which a set of disease nodes is linked with disease gene nodes (Goh & Choi, 2012). The objective of the construction of a

network for each complex disease holds the promise of identifying those interactions altered by mutations, which could help to find a treatment to revert the network back to normal state. The core of the human diseasome can be identified using a set of PPIs that are affected by a mutation leading to a pathological state. It can be obtained by purely computational tools and can help to highlight the key players that drive most of the characterized diseases (Janjic & Przulj, 2012). Even if the main disease-related proteins are identified, these advances do not mean a way to find a magic bullet for all pathologies. The highly dynamic nature of the signaling pathways due to their inter-connectivity is a characteristic that adds robustness to the cell (Kitano, 2004a). One example of a robust disease is cancer. A cancer tumor is a population of different cell types, each harboring their own mutations (Calon et al., 2012; Ding et al., 2012; Gerlinger et al., 2012; Hou et al., 2012). In this way, there are intracellular and intercellular interaction networks with different dynamics, since not all the proteome is expressed sequentially in a specific cue (S. P. Shah et al., 2012). Given the finite number of interactions between nodes in the cellular networks, there is a limit to the number of network configurations or states they can adopt. By rewiring the connections of a signaling network, cancer mutations are probably creating new states that are only present in cancer cells, and that are known as cancer network attractors states (Creixell, Schoof, Erler, & Linding, 2012).

The inter-connectivity of signaling pathways or pathway crosstalk is the underlying reason for such high network dynamics and is one of the reasons why a drug specifically designed for a key protein in a disease can fail. Thus, when a key pathway is inhibited, the cell may use another pathway that can have a similar physiological effect. The multiple layers of gene regulatory interactions modified by the alteration of the genetic material and structure (e.g. mutations in DNA, or aneuploidy at chromosomal level) combined with feedback loops give rise

to the robustness of the cancer cell. Thus, ‘de novo’ mutations during chemotherapy, in combination with feedback controls, allow the cancer cell to be resistant to treatment (Kitano, 2004b).

This is a problem from a pharmaceutical point of view, since a designed drug will be labeled as useless when it fails to stop the disease progression. Traditionally, the pharmacological approach to treat a disease has been a reductionist one, i.e. “one disease - one target - one drug”. In recent years, this has caused two major problems in the pharmaceutical field: 1) “me-too” drugs, when many companies design drugs for the same targets, and 2) poor assignment of medication to phenotypes due to multi-target properties (Frail & Barratt, 2012). The combination of systems biology with drug discovery, known as network pharmacology, is starting to change the approach of “one disease - one target - one drug” (Brown & Okuno, 2012). The generation of diseasome networks does not aim exclusively to determine the role of the gene or protein. We can add information such as the mutations that cause a given disease or confer susceptibility to a drug, in order to determine the role of individual players in the crosstalk context. A recent study showed that by using the pathway crosstalk data and available approved drugs it is possible to combine certain drugs targeting a particular signaling pathway in order to reduce the dose, while still being effective against cancer. As a consequence, this strategy has helped to develop an effective treatment less harmful to the patient (Jaeger & Aloy, 2012; Jaeger, Duran-Frigola, & Aloy, 2015).

Progress made with these different approaches has improved the rational design of drugs. Most of the designed drugs aim to block the binding sites of a protein. If the expected target of a drug is an enzyme, a first approach is to block the catalytic binding site, as in the case of neuraminidase inhibitors (Russell et al., 2006; Vavricka et al., 2011). An alternative approach to

target protein activity is by interfering protein interaction binding sites, therefore stabilizing or disrupting PPIs, like the transthyretin inhibitors (Sant'Anna et al., 2016). In fact, some mutations are lethal by modifying or interfering in a protein binding site, as in the case of the formation of amyloid fibrils that precedes the Amielod Lateral Sclerosis or Alzheimer's disease. In these cases, a mutation in the protein transthyretin destabilizes the formation of the normal multimer protein state, causing the proteins to aggregate in the form of fibrils. In this way, the mechanistic detail of how the protein is affected by drugs or mutations can only be given by the 3D structure of the protein and the complexes that it forms. Therefore, a high-quality image of the 3D structure of the proteins and the complexes they can form is an essential requirement for the design of effective drugs, which combined with the network approach, gives rise to new pharmacological strategies to treat disease in humans.

3. Structural approach to protein-protein interactions

Several diseases such as cancer or RASopaties (a group of diseases related to the malfunction of Ras signaling pathway), display altered PPIs networks (Kiel & Serrano, 2014). Current therapies that only target a single protein are not efficient in restoring the phenotype to normal in intricate signaling pathways. It would be needed to use a network-based therapeutic strategy to turn back the appearance of a malignant attractor state in the signaling network (Vidal et al., 2011). The use of pathway analysis on the network of interest could help to force the regression to the normal state. Current network maps give information on the relationships of genes or interactions between proteins (Figure 2). However, the vast majority of network analyses is done at a level of resolution that makes it difficult to include the three-dimensional (3D) structure of the cellular components at atomic level, a fundamental aspect that should be

taken into account (Kiel, Beltrao, & Serrano, 2008). From the amino acid sequence (primary structure), the inherent physicochemical properties of the polypeptide chain determine the first level of folding, known as secondary structure, with elements such as β-sheets or α-helices, as well as loops that do not fold into a specific structure. From this, combinations of β -sheets and α-helices can form the tertiary structure, where many proteins gain their functionality. The assembly of different polypeptide chains in complexes forms the quaternary structure. Databases such as STRING (von Mering et al., 2003) and Interactome3D (Mosca et al., 2013) provide curated information about the 3D structure of known protein-protein complexes. This type of information is of paramount importance for the understanding of biological processes at molecular level, as well as for applications in biomedicine such as rational drug design or repurposing studies, or interpretation of pathological mutations. Below are described the major experimental approaches to characterizing the structural details of protein-protein interactions.

[INSERT FIGURE 2 HERE]

3.1. X-ray Crystallography

The most widely used and accurate approach for obtaining high-resolution protein structures is the crystallography of proteins in combination with X-ray diffraction. A highly concentrated purified protein is needed for crystallization. Exposure of the crystal to an x-ray beam provides a diffraction spot pattern that gives information about “structures factors”, which allows building a map of electron density. The mathematical process to convert the intensities of the diffraction spots to the electron map is known as the phase resolution problem. The goal is to build a model of the protein based on this map, in which the protein sequence is the input to produce a

thermodynamically stable structure (Smyth & Martin, 2000). However, the process is very slow, requires a large amount of sample at a high purity quality, and often the protein has to be modified to achieve crystallization, with the risk of modifying the natural folding of the protein. Obtaining a crystal is not a routine process, since the conditions to find the formation of a crystal vary from sample to sample. Even after successfully obtaining a crystal, it might not be sufficiently optimal to determine the structure with high definition. Moreover, factors like the temperature and pH can affect the folding of the protein so that different structures can be obtained (Schiffer et al., 1989). In fact, there are cases where the applicability of this technique is extremely hard or unfeasible. Membrane proteins and low affinity complexes fall in this categorization since obtaining a crystal requires the stabilization by the membrane bilayer or a chemical scaffold to maintain the proteins folded and in close contact altering their natural conformation. Also, intrinsically disordered proteins, or very flexible loops present a problem since the periodicity required in for solving the phase problem cannot be easily achieved. Additionally, in some cases the use of a crystal structure as the representation of the biological relevant conformation of the protein in vivo has been challenged and is still under debate (Bahadur, Chakrabarti, Rodier, & Janin, 2004; Bahadur, Zacharias, & Janin, 2008; Ofran & Rost, 2007).

3.2. Nuclear Magnetic Resonance (NMR)

Another widely used technique to elucidate the 3D structure of a protein is Nuclear Magnetic Resonance (NMR). Since the 50’s NMR has evolved from the field of physics to the medical application. NMR relies on the use of strong magnetic fields where the nuclei and electrons of the atoms absorb the electromagnetic energy and reach a frequency of emission similar to the

natural isotopes (typically C13 and H1). However, this signal changes due the surrounding environment, thus giving also information of the nearby atoms. The advantage of NMR over the crystallography is that protein is in solution, a more natural environment that allows small movement of the proteins. It is very useful for determining the motions of proteins, including those large portions that do not have specific folding and are called intrinsically disordered. NMR experiments are time consuming and expensive, since larger molecules need machines with higher and higher frequency magnets. Thus, a major drawback of NMR is the size of the sample, since currently structures larger than 35 kDa cannot be determined. Therefore, in comparison with X-ray crystallography, very few complete structures of PPIs have been obtained by NMR, being especially difficult the case of multi complexes (Marion, 2013; N. Shah, Sattar, Benanti, Hollander, & Cheuck, 2006).

3.3. Cryogenic Electron Microscopy (Cryo-EM)

This technique is based on Electron Microscopy (EM). Standard EM needs to coat the sample with some special protector that usually contains metal particles like silver or gold, generating a layer with valleys and mountains according to the shape of the sample. Then, a laser is applied to the surface produced in the layer, creating the image in slices as it passes like in confocal microscopy. However, to enhance the image of minuscule samples, and to prevent degradation, and motion, the sample is fixed on a plate at very low temperatures, which is the basis for Cryo-EM.

Until recently Cryo-EM was regarded as a low-resolution technique because it presented a barrier at 6 Å of resolution and only allowed the inference of huge structures. However, with the recent improvement of the sensors, and high-level algorithms for image recognition, the

reconstruction of the 3D structure up to 2 Å resolution is possible (Elmlund & Elmlund, 2015). Still, many of the structures determined by this method are low-resolution and do not reveal the atomic details needed for most biological applications.

3.4. Small angle X-ray scattering

A recent structure determination development is the small angle X-ray scattering (SAXS). In contrast to crystallography, in SAXS the sample is exposed to an X-ray beam of a particular wavelength that is moved from 0 to 5 degrees to produce intensity distributions. The generated profile contains structural information of the atoms in the protein that can be in three different regions: the Guinier region that can be related to the average size of the group of atoms, the Fourier regions that contain information about the shape of the atoms in the protein, and the Porod region that provides information about the surface occupied in the volume by the atoms (Boldon, Laliberte, & Liu, 2015). The advantage of this method is that proteins can be studied in different media and even disordered. Interestingly, for the resolution of protein complexes, this technique can be coupled with other computational methods such as molecular dynamics or protein docking algorithms (Jimenez-Garcia, Pons, Svergun, Bernado, & Fernandez-Recio, 2015).

4. Computational modeling of protein-protein interactions

Despite all the recent advances, the majority of protein complexes are yet to be resolved. While there are 3D structures for nearly 50% of the proteins forming the human proteome (Muller, MacCallum, & Sternberg, 2002), only a small fraction (<7%) of the complexes forming the known human interactome are structurally characterized (Mosca et al., 2013). Thus, an option

to fill the structural gap in the human interactome is the use of computational modeling. The first attempt would be to construct the 3D structure of a complex from the amino acid sequence based on the available structure of complexes formed by similar proteins, using ab initio or homology-based modeling techniques similar to those used to model individual proteins. In this sense, the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction) (Kryshtafovych, Fidelis, & Moult, 2014) aims to assess how accurate is the prediction of current modeling programs in blind conditions. One approach is to make fully *ab initio* predictions from the protein sequences, considering the physicochemical properties of the amino acid and the energy terms that drive the folding. An alternative approach is to take advantage of the structures deposited in the PDB, by comparing gene products of different genes but with similar folding, so-called homology modeling. Homology modeling is a powerful tool to determine the 3D structure of proteins and complexes with a high degree of similarity. The most successful programs in CASP are multithreading software able to use structures deposited in the PDB, sequence similarity, and a little *ab initio* modeling. Winning strategies in the last editions of CASP are those of I-Tasser (Y. Zhang, 2008) and QUARK (Y. Zhang, 2014) which are programs that integrate fragment search in the PDB with the identification of basic folds that can be used as templates, and then fragments can be assembled into models of proteins.

4.1. Protein-protein docking

As above described, experimental determination of the structure of a PPI is highly challenging. Co-crystallizing two proteins is much more challenging than finding the right conditions for an individual protein; NMR has a size limitation, which leaves out mesoscopic protein ensembles, and Cryo-EM is still in development. As a consequence, all these experimental procedures can be

defined as low-throughput. These limitations create a gap between the number of new PPIs that are being discovered with high throughput experiments, and the very few 3D complex structures that are being determined. Computational approaches aim to compensate the difficulties in the determination of PPIs structures. However, predicting the 3D structure of the complex formed by two interacting proteins is a very challenging problem. The issue is similar to the structural prediction in individual structures, in the sense that both cases need a description of the physicochemical forces that regulate the interactions between the amino acid residues. Features such amino acid complementarity, electrostatics, steric clashes, hydrophobic effect, or hydrogen bonding, are concepts shared between both problems.

Unlike the problem of protein folding where the degrees of freedom in which a protein sequence can fold makes the space of search extremely large, in complex structure prediction, proteins are assumed to have 3D structure. This means that the search space is a six degree problem (three translations and three rotations), if we do not consider internal movements (rigid-body search). Computational tools such as protein-protein docking try to predict *ab initio* the correct orientation of two proteins that interact. Two major technical aspects can be found in the majority of docking methods: the generation of a large variety of structural models (sampling) and the identification of the correct docking poses with a proper function (scoring) (S. Y. Huang, 2014) (Figure 3). At the core of several docking protocols resides the idea of geometric complementarity in the protein-protein interface. However, in recent years different mechanisms have been proposed for protein-protein association: *i)* A basic mechanism called “lock and key”, directly inspired in complementarity, where the unbound monomers have a matching symmetry that is energetically favorable for the complex formation. This binding mechanism implies that both monomers are rigid, and they fit into one another; *ii)* The "induced fit" mechanism involves

conformational changes after binding of both monomers, before achieving the energetically favorable formed complex (Kuser, Cupri, Bleicher, & Polikarpov, 2008); and *iii)* The "conformational selection" mechanism assumes that bound states are naturally samples in the individual proteins and the binding partner selects those conformations that are energetically favorable for binding (Gianni, Dogan, & Jemth, 2014).

[INSERT FIGURE 3 HERE]

Protein-protein docking aims to predict the structure of a protein complex, inspired on the association mechanisms above described. In a real case scenario, the only information available is the 3D structure (or a reasonable model) of the unbound proteins. Current sampling strategies can be classified in: exhaustive global search, local shape feature matching, and randomized search.

Exhaustive global search over the protein aims to sample the entire possible space around a protein using as a probe another protein. In a rigid-body assumption, one needs to account for the translation on three axes, and the rotation on three axes, being a six degree of freedom problem. Exhaustive search can be achieved by using a grid to convert the surface of a protein into a coarse description. Then, Fast Fourier Transform (FFT) calculations (Katchalski-Katzir et al., 1992) can be used to reduce the computational cost by simplifying the translational and rotational search of the molecules. To completely search the 3D space of both proteins, one of the proteins (by convention the biggest one) is fixed and becomes the static molecule, while the other one moves in the 3D space through the FFT-based algorithm. The grid representation of the molecules allows to distinguish between the inside, the surface, and the outside of each protein.

The next step is to obtain a correlation score for all the relative translations between the two grids. This correlation score can be calculated on the molecular shape complementarity of the grids, by taking only into account the overlapping between surfaces as defined in the protein grids. After speeding the correlation calculation by FFT algorithms, a scoring function is applied and, this process repeats for each of the rotations of the mobile protein. This performs an exhaustive search of the 3D space of the interacting molecules. This method is by far the most popular one and has given rise to different programs where the differences are the description of the molecules on the grid. Some of these programs are FTDock (Gabb, Jackson, & Sternberg, 1997), ZDOCK (Chen, Li, & Weng, 2003), SDOCK (C. Zhang & Lai, 2011), PIPER (Kozakov, Brenke, Comeau, & Vajda, 2006), MolFit (Redington, 1992). A drawback of this type of approach is that it considers both proteins as rigid bodies, therefore, while it is suitable for an initial docking approach, it does not take into account the flexibility of both proteins. In fact, flexibility is one of the major current challenges for all docking algorithms.

Another approach is the local shape feature matching, with problem still remaining within six degrees of freedom. In this type of approach the molecular surface of both unbound proteins is calculated, which helps to identify binding regions. A segmentation algorithm is used to identify geometric features, such as convex, concave, and flat zones. Then, the molecular shape is represented by a graph in which each node is a representation for a surface region of the protein. The next step is to identify matching surfaces, with clique-search based approaches or geometric hashing. Programs like Patchdock (Schneidman-Duhovny, Inbar, Nussinov, & Wolfson, 2005), DOCK (Kuntz, Blaney, Oatley, Langridge, & Ferrin, 1982), or LZerD (Esquivel-Rodriguez, Filos-Gonzalez, Li, & Kihara, 2014) use this type of sampling to produce tens of thousands poses in a fast manner. One of the particular problems of this approach is that

the generated docking poses often include many atomic clashes, so additional steps of steric checking, clustering of solutions to avoid redundancy, or refinement are needed.

The third approach in sampling is random search. In this case, it is important to define several starting points and then drive the sampling towards the optimal positions. Some methods such as ICM disco (Fernandez-Recio, Totrov, & Abagyan, 2003), RossetaDock and HADDOCK (Dominguez, Boelens, & Bonvin, 2003) use random search as part of their docking strategy. Other algorithms are inspired by the swarms observed in the birds or insects, such as the Particle Swarm Optimization (PSO) (Clerc, 2010; Krishnanand & Ghose, 2008). For exploring the energetic landscape, the best energetic complexes are selected, and they are subsequently used as new seeds, with the process iterating until there are no new seeds. During the funnel-like search, the process only keeps the energetically favorable conformations and drives the docking proteins to the optimal matching pose. This type of algorithms can consider the flexibility of the proteins in the final refinement phase, during the minimization, or through normal mode representation of the search vectors. These algorithms are very successful to find near-native solutions, but computationally expensive. One successful example is the program SwarmDock (Moal & Bates, 2010).

4.2. Scoring of docking poses

Many current protein-protein docking protocols are successful if the interacting proteins undergo only small conformational changes upon binding. Even in these conditions, docking algorithms generate a large number of incorrect docking poses, so the aim is to place the near-native solutions as close to the top as possible within a ranked list. An important part of the success depends on the accuracy of the scoring function used to evaluate the docked conformations,

which in turns depends on its capabilities to overcome the inaccuracies of the interacting surfaces and singling out near-native conformations (Halperin, Ma, Wolfson, & Nussinov, 2002; Vajda & Kozakov, 2009). Generally speaking, scoring aims to identify the lowest-energy state among the different possible states of a given interaction, and thus, in the case of docking, it should be ideally able to describe the energetic aspects of protein-protein association (Moal & Fernandez-Recio, 2012). For practical predictions, the energy description of a system is estimated by approximate functions, and a large variety of scoring functions have been used, defined at different resolution levels (atomic or residue) (Tobi & Bahar, 2006). Docking algorithms often rely on the geometric complementarity of protein-protein interfaces. The essential zones for binding are often pre-formed in the interacting proteins (Levy, 2010), and as a consequence the interface of a protein complex could be considered an inherent geometric feature of the protein structures. This has made shape complementarity a popular ranking criterion to identify near-native solutions. Still, many protein-protein interfaces are flat, so complementarity alone is not enough to describe the right association mode. This is one of the reasons why a sampling step based only on geometry criteria often fails to produce correct models. Indeed, the physicochemical nature of the residues has a major role in protein association. Important elements include the electrostatic forces with complementary charges helping to provide the micro environment needed for the interface formation and the correct orientation of the proteins, and the hydrophobic effect with the burial of hydrophobic patches favoring the desolvation of the interacting surfaces (Camacho & Vajda, 2001; Camacho, Weng, Vajda, & DeLisi, 1999). Other factors are van der Waals attraction and repulsion, and hydrogen bonding. However, scoring functions that use energy-based terms to model these effects are not yet accurate enough to

reliably select near-native solutions from a pool of decoys, and thus further investigation is required to improve the quality of docking predictions.

Usually sampling and scoring are intimately coupled in a docking procedure. However, in many procedures, scoring is performed independently as a post-docking analysis. Basically, this approach consists in using a scoring function to re-rank the poses generated by a given docking program. This strategy could be considered as a type of refinement of the docking results, but using more sophisticated scoring functions than those used during the search phase. The idea behind post-docking approaches comes from the reasonable success of sampling algorithms to produce at least one near-native solution, also called a hit. In many cases, the in-built scoring function during the docking phase cannot be sensitive enough to place the near-native solution within the top of a ranked list of possible conformations. The computational problem is simplified by detaching the scoring functions from the sampling process, which also adds the possibility of combining different scoring functions. Some examples of post-docking methods are pyDock (Cheng, Blundell, & Fernandez-Recio, 2007), ZRANK (Pierce & Weng, 2007), SIPPER (Pons, Talavera, de la Cruz, Orozco, & Fernandez-Recio, 2011), DARS (Chuang et al., 2008). Given that docking programs typically report decoys ranked with only one or two scoring functions, it remains to be seen whether a given method could further benefit from the accumulated knowledge derived from the variety of currently available scoring functions that have been reported in the literature, many of which were developed for different modeling problems (Tobi, 2010). One example of this is the combination PIE/PIER (Viswanath, Ravikant, & Elber, 2013). In some methods, the scoring functions are also combined with the inclusion of protein flexibility, like in Fiberdock (Mashiach, Nussinov, & Wolfson, 2010), Firedock (Andrusier, Nussinov, & Wolfson, 2007), or RDOCK (L. Li, Chen, & Weng, 2003).

Among the different scoring functions applied as post-docking analysis, we note the program pyDock (Cheng et al., 2007), which is a well-known protein-protein protocol using the FTDock or ZDOCK sampling combined with a highly efficient scoring function. The pyDock scoring function is formed by three energy-based terms: Coulombic electrostatics, desolvation energy and van der Waals potential. A protein is a charged entity and its surface has to be in constant contact with solvent molecules, so considering the electrostatic charges of the proteins is the basis of the majority of the scoring functions. But electrostatics alone is not enough to place the two interacting proteins in the optimal position, so there is a need for additional terms to help to improve the algorithm. Since many of the binding surfaces are flat, and the critical contact residues at the interface are often hydrophobic, desolvation plays a major role in creating the micro-environment necessary to allow the formation of a strong interaction between proteins. On the other side, the van der Waals energy is usually important for the final assembly of two given proteins, and it is very dependent on the correct side-chain conformations. When docking is rigid-body, this potential is very noisy. The use of all the above mentioned energy descriptors makes pyDock a very versatile, non-deterministic, and adaptable docking method.

4.3. Template-based docking

In addition to ab initio docking, the interface between two interacting proteins could be modeled using the existing structural data in the PDB (Sinha, Kundrotas, & Vakser, 2012). Figure 4 shows a schematic view of a template-based docking protocol in comparison with *ab initio* docking. As seen in modeling of individual proteins, some evolutionary distant PPIs converged in a structural conformation which is optimal for the recognition (interologs) (Matthews et al., 2001). The identification of interologs facilitates the study of PPIs. The conservation of the structural

conformation of the interface through evolution has also demonstrated a plasticity to changes, where only 66% of the interface patch is conserved, leaving the remaining 34% of the interface tolerant to residue changes (Faure, Andreani, & Guerois, 2012). However, the interface is also a dynamic part of the protein that can change during binding (Hamp & Rost, 2012). In fact, the inclusion of evolutionary data in the context of interface predictions seem to give additional confidence in the prediction (Hamp & Rost, 2015; Katsonis & Lichtarge, 2014).

[INSERT FIGURE 4 HERE]

It has been recently claimed that there could be available templates for most of the known protein complexes (Kundrotas, Zhu, Janin, & Vakser, 2012). However, in the case of remote homology, i.e. the twilight zone, the available templates do not provide better modeling than *ab initio* docking (Negroni, Mosca, & Aloy, 2014).

4.4. Interface and hot-spot prediction

The use of new approaches continues to enable the study of protein interactions from different perspectives. The analysis of protein-protein complex structures have aimed to identify different properties that can distinguish protein-protein interfaces from the rest of the protein surface (Jones & Thornton, 1997). The protein-protein interface is a critical zone for molecular recognition, formed by an average of ~28 residues, accounting for around 1000 A^2 of the area in one protein, and mostly flat. Based on the relative Accessible Surface Area (rASA) of the residues in the interface, three different zones could be defined (Levy, 2010), as shown in Figure 5: *i)* core, formed by residues that are exposed in the unbound monomers (rASA unbound >

25%) and become buried in the complex (rASA complex < 25%), forming the necessary contacts for the interaction and contributing largely to the binding energy; *ii)* rim residues, which are exposed in the unbound monomers (rASA unbound > 25%) and, although to a lesser extent, remain exposed in the complex (rASA complex > 25%), shielding the core from the solvent and providing the micro-environment required for establishing the interaction; and *iii)* support, formed by residues that are largely buried in the unbound monomers (rASA unbound < 25%), and become more buried in the complex (rASA complex < 25%), helping to establish the interaction.

[INSERT FIGURE 5 HERE]

Interface residues seem to play different roles in disease according to the region they belong to. In a recent study, it was found that the core interface residues are more susceptible to disease-related mutations, in contrast to those in the rim regions (David & Sternberg, 2015). Complementary work showed that about 11% of all known disease-associated SNPs also land outside but near to the interface (Gao, Zhou, & Skolnick, 2015). Both studies found that the residues that are more vulnerable to disease-related mutations are residues buried in the interface, although they seem to differ about the preferred localization of these mutations.

Alanine scanning (Morrison & Weiss, 2001) can be used to experimentally describe the contribution of the different residues to the interaction. The technique consists in performing point mutations in the protein sequence for alanine, so that the chemical neutral nature and size of the alanine residue mimics the removal of a given residue without perturbing too much the secondary structure. Based on this technique, experimental analyses have shown that most of the

binding affinity is contributed by just a small number of interface residues, called *hot-spots*, which are often found at the interface core (Clackson & Wells, 1995). The identification of such hot-spot residues at protein-protein interfaces in complexes of biomedical interest is relevant for drug discovery purposes, as they are suitable targets for small-molecules capable of modulating the interaction. However, experimental determination of hot-spots by alanine scanning is costly and time consuming. This has fostered the development of many computational approaches that aim to complement experimental data. The vast majority of the predictive methods strongly rely on the availability of the complex structure. Several energy-based methods have been reported, such as ROBETTA (Kortemme & Baker, 2002), FoldX (Schymkowitz et al., 2005), HSPred (Lise, Buchan, Pontil, & Jones, 2011) or Molecular Dynamics (MD) with generalized Born model in a continuum medium (Moreira, Fernandes, & Ramos, 2007), supported in several MD platforms (e.g., AMBER (Salomon-Ferrer, Case, & Walker, 2013) and GROMACS (Pronk et al., 2013)), which are based on computational alanine scanning of protein-protein interfaces and subsequent evaluation of the change in binding affinity.

Other valuable approaches are based on machine learning. Recently reported methods are KFC2 (Zhu & Mitchell, 2011), based on interface solvation, atomic density and plasticity features; PCRPi (Assi, Tanaka, Rabbitts, & Fernandez-Fuentes, 2010), combining sequence conservation, energy score and contact number information; PPI-Pred (Bradford & Westhead, 2005), considering surface shape and electrostatics; MINERVA, which weights atomic packing density and hydrophobicity (K. I. Cho, Kim, & Lee, 2009) or a neural network-based protocol (an adaptation of ISIS), which combines several interface features such as sequence profiles, solvent accessibility and evolutionary conservation (Ofran & Rost, 2007). Another well-known machine learning-based tool is PocketQuery web-server (Koes & Camacho, 2012), which

provides an assortment of metrics (including changes in solvent accessible surface area, energy-based scores, and sequence conservation) extremely useful for hot-spots, anchor residues and hot regions prediction.

Empirical formula-based methods are also used instead of machine learning algorithms, such as MAPPIS (Shulman-Peleg, Shatsky, Nussinov, & Wolfson, 2007), which relies on the evolutionary conservation of hot-spots in the interface along different family members; HotSpot Wizard (Pavelka, Chovancova, & Damborsky, 2009), based on the integration of structural, functional and evolutionary information provided by several databases; DrugScorePPI (Kruger, Ignacio Garzon, Chacon, & Gohlke, 2014), derived from experimental alanine scanning results; iPRED (Geppert, Hoy, Wessler, & Schneider, 2011), using pairwise potential atom types and residue properties; APIS (Xia, Zhao, Song, & Huang, 2010), where the hot-spots identification is performed by combining residue physical/biochemical features, such as protrusion index and solvent accessibility; HotPoint (Tuncbag, Keskin, & Gursoy, 2010), using occlusion from solvent and knowledge-based pair residue potentials; and ECMIS (Shingate, Manoharan, Sukhwal, & Sowdhamini, 2014), using a new algorithm combining energetic, evolutionary and structural features.

In spite of their high accuracy in the identification of hot-spot residues, a major limitation of all the above cited tools is that they depend on the availability of the protein-protein complex structure (or a reliable model). However, for the majority of interactions, the complex structure is not available, and as a consequence these tools cannot be used. A very few hot-spot prediction methods have been reported that do not need the structure of the complex. One of such methods is pyDockNIP (Grosdidier & Fernandez-Recio, 2008), which is based on the analysis of protein-protein docking models generated with pyDock (Cheng et al., 2007). The method computes the

propensity of a given residue to be located at the interface in the 100 lowest-energy rigid body docking solutions, and can reach high precision in the prediction of hot-spots, but at the expense of low sensitivity. Another method that do not need the complex structure is SIM (Agrawal, Helk, & Trout, 2014), which predicts hot-spot residues involved in evolutionarily conserved protein-protein interactions.

All the different methods for computational analysis and prediction of interface and hot-spot residues have inspired the creation of several databases of computationally predicted hot-spot residues, such as HotRegion (Cukuroglu et al., 2014), HotSprint (Guney, Tuncbag, Keskin, & Gursoy, 2008) and PCRPi-DB (Segura & Fernandez-Fuentes, 2011).

4.5. Assessment of protein-protein docking predictions

In order to assess the predictive accuracy of a newly developed method, it would be necessary to have a reference set widely accepted by the community. In the case of the protein-protein docking field, the reference set needs to have the crystal structure of the proteins in a free state and that of the complexed or bound state. These structures must have a high resolution, and good coverage of the proteins. In addition, the protein set should to be diverse enough, so that it can represent as many as the known protein families as possible. The current version of the most widely used protein-protein docking benchmark has 231 protein complexes (Vreven et al., 2015). Each of those complexes has the crystal structure of the proteins in unbound form and the bound form. The protein docking benchmark is divided into subcategories according to the difficulty, based on the conformational changes that the proteins undergo from unbound to bound states. The most difficult category corresponds to the cases that are the most difficult to predict with

current protein docking algorithms, mostly due to the large conformational changes of the proteins.

Other benchmarks have been reported to assess different methods for PPI modeling, like binding affinity changes upon mutation (Moal & Fernandez-Recio, 2012), scorer sets from CAPRI (Lensink & Wodak, 2014), or binding affinity data sets (Kastritis et al., 2011). There are other useful databases such as template libraries (DOCKGROUND (Liu, Gao, & Vakser, 2008)), structural datasets with similarity between sequences (3D-Complex (Levy et al., 2006)), or classification of the domain-domain interaction on protein complexes like SCOPPI (Winter et al., 2006).

Protein-protein docking programs are blindly assessed in the Critical Assessment of PRedicted Interactions (CAPRI) (Janin et al., 2003), which is an international scientific effort to boost the development of different approaches to solve the problem of protein-protein docking. After more than fifteen years since the first edition, the CAPRI experiment is now the source of standard protein-protein docking sets and quality measurements.

5. Application of computational docking to biomedicine

5.1. Interpretation of pathological mutations perturbing protein-protein interactions

As above mentioned, it would be important to estimate the effects of a given gene variant at molecular level, which will contribute to understand better the phenotype related to such variant, e.g. pathological condition, disease predisposition, altered drug response, etc. as well as to rationalize therapeutic intervention, within the context of personalized medicine. For this, it is essential to understand the role of the protein interaction networks in a particular biological process, and how genetic variants such as non-synonymous single nucleotide polymorphisms

(nsSNPs) can affect specific protein-protein interactions (Rual et al., 2005; Wu et al., 2014). When a mutation has a strong effect on the folding or stability of a protein, it may disrupt all interactions of the mutated protein. However, if the mutation is located at a specific protein-binding interface, it could affect only some of the interactions of the mutated protein or "edges" in a particular network (so called edgetic effect) (David & Sternberg, 2015; Sahni et al., 2015; Zanzoni, Soler-Lopez, & Aloy, 2009; Zhong et al., 2009). In each of these situations, the observed disorders are ultimately different, as well as their causes, consequences, and therapeutic options (Figure 6).

[INSERT FIGURE 6 HERE]

Indeed, large-scale structural analyses show that pathological mutations are enriched on the domains that are relevant for protein-protein interactions (Wang et al., 2012) and many disease-related mutations are directly involved at protein-protein interfaces (David, Razali, Wass, & Sternberg, 2012; David & Sternberg, 2015; Mosca et al., 2015). It has been found that missense mutations described in the database OMIM can cause changes in protein-protein binding energy (Teng, Madej, Panchenko, & Alexov, 2009). The integration of structural data in proteins complexes with interaction network description can help to understand the effect of disease-related mutations at molecular level (Fraser, Gross, & Krogan, 2013). Through a combination of interaction network analysis, structural data and energetic calculations, many of the known pathological mutations involved in cancer and/or RASopathies have been found to have a direct effect on the binding affinity in some of the interactions in the RAS/MAPK cascade (Kiel & Serrano, 2012, 2014). Moreover, this effect (together with other structural and energetics

effects) can provide a first general explanation for some of the differences in phenotype. More recently, interaction perturbation profiling of missense mutations across a broad spectrum of human disorders suggests that around one third of disease mutations have edgetic effects, that is, they only affect to specific interactions of a given protein, as opposed to structural mutations that can perturb simultaneously all the interactions (Sahni et al., 2015). Interestingly, mutated proteins with edgetic effects have been found to play central roles in the protein network (Sahni et al., 2015). This is a direct explanation at molecular level of how dissimilar mutations within the same gene may produce distinct interaction profiles and, as a consequence, different disease phenotypes (Sahni et al., 2015).

Understanding the role of pathological mutations in protein-protein interactions can help to improve our knowledge of disease at molecular level, which could be very important for predicting pathogenicity in missense mutations. The functional prediction of mutations has a growing importance in clinical practice, especially when dealing with patient mutations that are not annotated or that have unclear diagnosis, prognosis or disease development. In these situations, general pathogenicity prediction methods are used, such as PolyPhen-2 (Adzhubei, Jordan, & Sunyaev, 2013; Adzhubei et al., 2010), SIFT (Sim et al., 2012), or PON-P2 (Niroula, Urolagin, & Vihinen, 2015), which can help physicians to make clinical decisions. These methods have good prediction rates in general, but they fail in many specific diseases (Riera, Padilla, & de la Cruz, 2016). Indeed, current models cannot correctly describe all the effects that amino acid mutations can cause in proteins, such as to what extent the mutation is perturbing the interactions to other proteins and biomolecules (Cheng et al., 2012), as above mentioned.

Thus, for a more complete characterization of a given mutation at molecular level, with pathogenicity prediction purposes, it would be needed: i) to structurally characterize the location

of the mutation with respect to protein-binding interfaces, and ii) to characterize the energetic effect on the binding affinity of the involved interactions. However, one of the big limitations in the field is the small number of protein-protein complexes with their 3D structure deposited in the Protein Data Bank (PDB) (Berman et al., 2002). The structures of weak or transient complexes, dynamic assemblies, or multi-protein associations are particularly difficult to determine by crystallography or NMR, as above mentioned. As a consequence, there is a growing gap between the number of protein complexes with available experimental structure and the number of interactions that are being discovered. While around half of the non-redundant proteins in human have available structure (or a reliable model), less than 7% of the estimated number of protein-protein interactions in human have available structure (Mosca et al., 2013). In this context, computational docking methods (Chen et al., 2003; Cheng et al., 2007; Mashiach et al., 2010; C. Zhang & Lai, 2011) are already being used to model the structure of protein-protein complexes of biomedical interest, so in principle, they could be very useful for the structural characterization of entire interactomes (Mosca, Pons, Fernandez-Recio, & Aloy, 2009). However, for many cases, structural prediction by docking is not accurate enough, so its application at interactomic scale is not yet practical. More accurate is the prediction of interface residues, usually based on sequence conservation or physicochemical properties, which can be more useful for large-scale analyses. But not all the interface residues contribute equally to binding affinity. Thus, it is important to identify the so-called "hot-spot" residues, which are those that contribute the most to the binding energy (Clackson & Wells, 1995). Previous work showed that these interface hot-spot residues can be identified based on docking calculations, even if the structure of the protein-protein complex is not available (Grosdidier & Fernandez-Recio, 2008).

The combination of structural information and docking-based modeling will be essential for the interpretation of pathological mutations at interactomic level.

5.2. Chemical perturbation of disease-related protein-protein interactions for drug discovery

As above mentioned, pathological mutations can significantly perturb specific protein-protein interactions, either by disrupting these interactions or by stabilizing them (Rolland et al., 2014). In either case, such perturbed interactions constitute an attractive target for therapeutic intervention. Indeed, the identification of modulators of specific protein-protein interactions (e.g. PPI inhibitors) is the next milestone in the drug discovery field (Wells & McClendon, 2007). Several examples of PPI peptide inhibitors have been reported based on mimetic peptides that replace the interaction surface of one of the proteins. But the lower bioavailability of peptides makes them to be not very attractive for therapeutic purposes. Therefore, it is necessary to apply structure-based approaches to identify small molecules capable of inhibiting PPI. However, protein–protein interactions differ from traditional drug target proteins in that: i) protein-protein interfaces are large and involve more atomic interactions and hence higher affinity as compared to protein-ligand interfaces; ii) protein-protein interfaces do not have clear binding pockets as in the case of traditional protein drug targets; and iii) most often, the location of the interface and the binding mode of the targeted interaction is not known. All of the above considerations pose clear difficulties to apply standard drug discovery procedures.

The first difficulty is that protein-protein interfaces (PPIs) are much larger (~1500-3000 $Å^2$) than protein-small molecule interfaces (~300-1000 $Å^2$), which makes it difficult to find small molecules to disrupt PPIs. We have mentioned above the existence of hot-spot residues, which

are important in the context of drug discovery because targeting them is the only way for a small-molecule to compete with a protein-protein interaction. Although there are available experimental data about hot-spot residues for a few complexes, it is necessary to complement the costly experimental procedures with computational approaches.

The second difficulty is the absence of natural pockets in protein-protein interfaces. It is necessary to describe the possible fluctuations of the interacting molecules in order to find transient pockets that can be useful for drug discovery (Eyrisch & Helms, 2007).

Last but not less important is the absence of structural information for the majority of protein-protein interactions. When there is no available structure for the complex, it is necessary to know at least the location of the protein-protein interface is in order to narrow the search for transient pockets suitable for small-molecule docking (Figure 7).

[INSERT FIGURE 7 HERE]

In order to help solving all the above mentioned difficulties to identify modulators of protein-protein interactions, computational approaches such as protein-protein docking and molecular dynamics are becoming increasingly important tools in drug discovery. Protein-protein docking aims to predict the structure of a protein-protein complex starting from the 3-D coordinates of the unbound structures. As mentioned in previous sections, the docking program pyDock (Cheng et al., 2007) can be applied to predict protein interfaces and to identify the most relevant residues in protein-protein interactions (hot-spots) when there is no structural information about the protein-protein complex (Grosdidier & Fernandez-Recio, 2008, 2012). Molecular dynamics (MD) is another computational approach that can be also applied to find

possible transient pockets within protein-protein interfaces, together with computational tools capable of identifying suitable cavities in the protein surfaces, such as Fpocket (Le Guilloux, Schmidtke, & Tuffery, 2009), QsiteFinder (Laurie & Jackson, 2005), PASS (Brady & Stouten, 2000) and LigSite (Hendlich, Rippmann, & Barnickel, 1997).

REFERENCES

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet, Chapter 7*, Unit7 20. doi:10.1002/0471142905.hg0720s76

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Agrawal, N. J., Helk, B., & Trout, B. L. (2014). A computational tool to predict the evolutionarily conserved protein-protein interaction hot-spot residues from the structure of the unbound protein. *FEBS Lett, 588*(2), 326-333. doi:10.1016/j.febslet.2013.11.004

Albert, R., Jeong, H., & Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature, 406*(6794), 378-382. doi:10.1038/35019019

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Andrusier, N., Nussinov, R., & Wolfson, H. J. (2007). FireDock: fast interaction refinement in molecular docking. *Proteins, 69*(1), 139-159. doi:10.1002/prot.21495

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., . . . Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res, 38*(Database issue), D525-531. doi:10.1093/nar/gkp878

Arnau, V., Mars, S., & Marin, I. (2005). Iterative cluster analysis of protein interaction data. *Bioinformatics, 21*(3), 364-378. doi:10.1093/bioinformatics/bti021

Assi, S. A., Tanaka, T., Rabbitts, T. H., & Fernandez-Fuentes, N. (2010). PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res, 38*(6), e86. doi:10.1093/nar/gkp1158

Bahadur, R. P., Chakrabarti, P., Rodier, F., & Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol, 336*(4), 943-955.

Bahadur, R. P., Zacharias, M., & Janin, J. (2008). Dissecting protein-RNA recognition sites. *Nucleic Acids Res, 36*(8), 2705-2716. doi:10.1093/nar/gkn102

Baoying, W., Ruowang, L., & William, P. (Eds.). (2015). *Big Data Analytics in Bioinformatics and Healthcare*. Hershey, PA, USA: IGI Global.

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet, 5*(2), 101-113. doi:10.1038/nrg1272

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2005). GenBank. *Nucleic Acids Res, 33*(Database issue), D34-38. doi:10.1093/nar/gki063

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., . . . Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr, 58*(Pt 6 No 1), 899-907.

Bhalla, U. S., & Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science*, *283*(5400), 381-387.

Bharathy, N., & Taneja, R. (2012). Methylation muscles into transcription factor silencing. *Transcription, 3*(5), 215-220. doi:10.4161/trns.20914

Boldon, L., Laliberte, F., & Liu, L. (2015). Review of the fundamental theories behind small angle X-ray scattering, molecular dynamics simulations, and relevant integrated application. *Nano Rev, 6*, 25661. doi:10.3402/nano.v6.25661

Bradford, J. R., & Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics, 21*(8), 1487-1494. doi:10.1093/bioinformatics/bti242

Brady, G. P., Jr., & Stouten, P. F. (2000). Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des, 14*(4), 383-401.

Brown, J. B., & Okuno, Y. (2012). Systems biology and systems chemistry: new directions for drug discovery. *Chem Biol, 19*(1), 23-28. doi:10.1016/j.chembiol.2011.12.012

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1002822

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. V., Iglesias, M., Cespedes, M. V., . . . Batlle, E. (2012). Dependency of colorectal cancer on a TGF-beta-driven program in stromal cells for metastasis initiation. *Cancer Cell, 22*(5), 571-584. doi:10.1016/j.ccr.2012.08.013

Camacho, C. J., & Vajda, S. (2001). Protein docking along smooth association pathways. *Proc Natl Acad Sci U S A, 98*(19), 10636-10641. doi:10.1073/pnas.181147798

Camacho, C. J., Weng, Z., Vajda, S., & DeLisi, C. (1999). Free energy landscapes of encounter complexes in protein-protein association. *Biophys J, 76*(3), 1166-1178. doi:10.1016/S0006-3495(99)77281-4

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., . . . Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res, 45*(D1), D369-D379. doi:10.1093/nar/gkw1102

Chen, R., Li, L., & Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins, 52*(1), 80-87. doi:10.1002/prot.10389

Cheng, T. M., Blundell, T. L., & Fernandez-Recio, J. (2007). pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins, 68*(2), 503-515. doi:10.1002/prot.21419

Cheng, T. M., Goehring, L., Jeffery, L., Lu, Y. E., Hayles, J., Novak, B., & Bates, P. A. (2012). A structural systems biology approach for quantifying the systemic consequences of missense mutations in proteins. *PLoS Comput Biol, 8*(10), e1002738. doi:10.1371/journal.pcbi.1002738

Cho, K. I., Kim, D., & Lee, D. (2009). A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res, 37*(8), 2672-2687. doi:10.1093/nar/gkp132

Cho, Y. R., Hwang, W., Ramanathan, M., & Zhang, A. (2007). Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics, 8*, 265. doi:10.1186/1471-2105-8-265

Chowbina, S. R., Wu, X., Zhang, F., Li, P. M., Pandey, R., Kasamsetty, H. N., & Chen, J. Y. (2009). HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics, 10 Suppl 11*, S5. doi:10.1186/1471-2105-10-S11-S5

Chuang, G. Y., Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2008). DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophys J, 95*(9), 4217-4227. doi:10.1529/biophysj.108.135814

Clackson, T., & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science, 267*(5196), 383-386. doi:10.1126/science.7529940

Clerc, M. (2010). Conclusion *Particle Swarm Optimization* (pp. 189-191): ISTE.

Cooper, G. M., Johnson, J. A., Langaee, T. Y., Feng, H., Stanaway, I. B., Schwarz, U. I., . . . Rieder, M. J. (2008). A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood, 112*(4), 1022-1027. doi:10.1182/blood-2008-01-134247

Creixell, P., Schoof, E. M., Erler, J. T., & Linding, R. (2012). Navigating cancer network attractors for tumor-specific therapy. *Nat Biotechnol, 30*(9), 842-848. doi:10.1038/nbt.2345

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*(5258), 561-563.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., . . . Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res, 39*(Database issue), D691-697. doi:10.1093/nar/gkq1018

Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y., . . . Shi, T. (2008). AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology. *Nucleic Acids Res, 36*(Database issue), D999-1008. doi:10.1093/nar/gkm844

Cukuroglu, E., Gursoy, A., Nussinov, R., & Keskin, O. (2014). Non-redundant unique interface structures as templates for modeling protein interactions. *PLoS One, 9*(1), e86738. doi:10.1371/journal.pone.0086738

David, A., Razali, R., Wass, M. N., & Sternberg, M. J. (2012). Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum Mutat, 33*(2), 359-363. doi:10.1002/humu.21656

David, A., & Sternberg, M. J. (2015). The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease. *J Mol Biol, 427*(17), 2886-2898. doi:10.1016/j.jmb.2015.07.004

Derkatch, I. L., & Liebman, S. W. (2007). Prion-prion interactions. *Prion, 1*(3), 161-169.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., . . . DiPersio, J. F. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature, 481*(7382), 506-510. doi:10.1038/nature10738

Dominguez, C., Boelens, R., & Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc, 125*(7), 1731-1737. doi:10.1021/ja026939x

Elmlund, D., & Elmlund, H. (2015). Cryogenic electron microscopy and single-particle analysis. *Annu Rev Biochem, 84*, 499-517. doi:10.1146/annurev-biochem-060614-034226

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res, 30*(7), 1575-1584.

Esquivel-Rodriguez, J., Filos-Gonzalez, V., Li, B., & Kihara, D. (2014). Pairwise and multimeric protein-protein docking using the LZerD program suite. *Methods Mol Biol, 1137*, 209-234. doi:10.1007/978-1-4939-0366-5_15

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., . . . Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol, 3*, 89. doi:10.1038/msb4100134

Eyrisch, S., & Helms, V. (2007). Transient pockets on protein surfaces involved in protein-protein interaction. *J Med Chem, 50*(15), 3457-3464. doi:10.1021/jm070095g

Faure, G., Andreani, J., & Guerois, R. (2012). InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res, 40*(Database issue), D847-856. doi:10.1093/nar/gkr845

Fernandez-Recio, J., Totrov, M., & Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins, 52*(1), 113-117. doi:10.1002/prot.10383

Fields, S., & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature, 340*(6230), 245-246. doi:10.1038/340245a0

Frail, D. E., & Barratt, M. J. (2012). Opportunities and Challenges Associated with Developing Additional Indications for Clinical Development Candidates and Marketed Drugs *Drug Repositioning* (pp. 33-51): John Wiley & Sons, Inc.

Fraser, J. S., Gross, J. D., & Krogan, N. J. (2013). From systems to structure: bridging networks and mechanism. *Molecular cell, 49*(2), 222-231. doi:10.1016/j.molcel.2013.01.003

Freedman, M. L., Monteiro, A. N., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C., . . . Mills, I. G. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet, 43*(6), 513-518. doi:10.1038/ng.840

Gabb, H. A., Jackson, R. M., & Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol, 272*(1), 106-120. doi:10.1006/jmbi.1997.1203

Gao, M., Zhou, H., & Skolnick, J. (2015). Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure, 23*(7), 1362-1369. doi:10.1016/j.str.2015.03.028

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. doi:10.1038/nature09534

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Geppert, T., Hoy, B., Wessler, S., & Schneider, G. (2011). Context-based identification of protein-protein interfaces and "hot-spot" residues. *Chem Biol, 18*(3), 344-353. doi:10.1016/j.chembiol.2011.01.005

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., . . . Swanton, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med, 366*(10), 883-892. doi:10.1056/NEJMoa1113205

Gianni, S., Dogan, J., & Jemth, P. (2014). Distinguishing induced fit from conformational selection. *Biophys Chem, 189*, 33-39. doi:10.1016/j.bpc.2014.03.003

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., . . . Rothberg, J. M. (2003). A protein interaction map of Drosophila melanogaster. *Science, 302*(5651), 1727-1736. doi:10.1126/science.1090289

Glazko, G. V., & Emmert-Streib, F. (2009). Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics, 25*(18), 2348-2354. doi:10.1093/bioinformatics/btp406

Goh, K. I., & Choi, I. G. (2012). Exploring the human diseasome: the human disease network. *Brief Funct Genomics, 11*(6), 533-542. doi:10.1093/bfgp/els032

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabasi, A. L. (2007). The human disease network. *Proc Natl Acad Sci U S A, 104*(21), 8685-8690. doi:10.1073/pnas.0701361104

Grosdidier, S., & Fernandez-Recio, J. (2008). Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics, 9*, 447. doi:10.1186/1471-2105-9-447

Grosdidier, S., & Fernandez-Recio, J. (2012). Protein-protein docking and hot-spot prediction for drug discovery. *Curr Pharm Des, 18*(30), 4607-4618.

Guney, E., Tuncbag, N., Keskin, O., & Gursoy, A. (2008). HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res, 36*(Database issue), D662-666. doi:10.1093/nar/gkm813

Guruharsha, K. G., Rual, J. F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., . . . Artavanis-Tsakonas, S. (2011). A protein complex network of Drosophila melanogaster. *Cell, 147*(3), 690-703. doi:10.1016/j.cell.2011.08.047

Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins, 47*(4), 409-443. doi:10.1002/prot.10115

Hamp, T., & Rost, B. (2012). Alternative protein-protein interfaces are frequent exceptions. *PLoS Comput Biol, 8*(8), e1002623. doi:10.1371/journal.pcbi.1002623

Hamp, T., & Rost, B. (2015). Evolutionary profiles improve protein-protein interaction prediction from sequence. *Bioinformatics, 31*(12), 1945-1950. doi:10.1093/bioinformatics/btv077

Hartuv, E., & Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters, 76*(4-6), 175-181. doi:10.1016/s0020-0190(00)00142-3

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature, 402*(6761 Suppl), C47-52. doi:10.1038/35011540

Hendlich, M., Rippmann, F., & Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model, 15*(6), 359-363, 389.

Higgins, D. G., & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene, 73*(1), 237-244. doi:https://doi.org/10.1016/0378-1119(88)90330-7

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., . . . Wang, J. (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell, 148*(5), 873-885. doi:10.1016/j.cell.2012.02.028

Huang, H., Wu, X., Sonachalam, M., Mandape, S. N., Pandey, R., MacDorman, K. F., . . . Chen, J. Y. (2012). PAGED: a pathway and gene-set enrichment database to enable molecular

phenotype discoveries. *BMC Bioinformatics, 13 Suppl 15*, S2. doi:10.1186/1471-2105-13-S15-S2

Huang, S. Y. (2014). Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discov Today, 19*(8), 1081-1096. doi:10.1016/j.drudis.2014.02.005

Human Genome Sequencing, C. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931-945. doi:http://www.nature.com/nature/journal/v431/n7011/suppinfo/nature03001_S1.html

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., . . . Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science, 292*(5518), 929-934. doi:10.1126/science.292.5518.929

International HapMap, C., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., . . . McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature, 467*(7311), 52-58. doi:10.1038/nature09298

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., . . . Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A, 97*(3), 1143-1147.

Jaeger, S., & Aloy, P. (2012). From protein interaction networks to novel therapeutic strategies. *IUBMB Life, 64*(6), 529-537. doi:10.1002/iub.1040

Jaeger, S., Duran-Frigola, M., & Aloy, P. (2015). Drug sensitivity in cancer cell lines is not tissue-specific. *Mol Cancer, 14*, 40. doi:10.1186/s12943-015-0312-6

Janes, K. A., Albeck, J. G., Gaudet, S., Sorger, P. K., Lauffenburger, D. A., & Yaffe, M. B. (2005). A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science, 310*(5754), 1646-1653. doi:10.1126/science.1116598

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., . . . Critical Assessment of, P. I. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins, 52*(1), 2-9. doi:10.1002/prot.10381

Janjic, V., & Przulj, N. (2012). The Core Diseasome. *Mol Biosyst, 8*(10), 2614-2625. doi:10.1039/c2mb25230a

Jimenez-Garcia, B., Pons, C., Svergun, D. I., Bernado, P., & Fernandez-Recio, J. (2015). pyDockSAXS: protein-protein complex structure by SAXS and computational docking. *Nucleic Acids Res, 43*(W1), W356-361. doi:10.1093/nar/gkv368

Jones, S., & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol, 272*(1), 121-132. doi:10.1006/jmbi.1997.1234

Kandpal, R., Saviola, B., & Felton, J. (2009). The era of 'omics unlimited. *Biotechniques, 46*(5), 351-352, 354-355. doi:10.2144/000113137

Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M., & Janin, J. (2011). A structure-based benchmark for protein-protein binding affinity. *Protein Sci, 20*(3), 482-491. doi:10.1002/pro.580

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., & Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A, 89*(6), 2195-2199.

Katsonis, P., & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Res, 24*(12), 2050-2058. doi:10.1101/gr.176214.114

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol, 8*(2), e1002375. doi:10.1371/journal.pcbi.1002375

Kiel, C., Beltrao, P., & Serrano, L. (2008). Analyzing protein interaction networks using structural information. *Annu Rev Biochem, 77*, 415-441. doi:10.1146/annurev.biochem.77.062706.133317

Kiel, C., & Serrano, L. (2012). Structural data in synthetic biology approaches for studying general design principles of cellular signaling networks. *Structure, 20*(11), 1806-1813. doi:10.1016/j.str.2012.10.002

Kiel, C., & Serrano, L. (2014). Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. *Mol Syst Biol, 10*, 727. doi:10.1002/msb.20145092

King, A. D., Przulj, N., & Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics, 20*(17), 3013-3020. doi:10.1093/bioinformatics/bth351

Kitano, H. (2004a). Biological robustness. *Nat Rev Genet, 5*(11), 826-837. doi:10.1038/nrg1471

Kitano, H. (2004b). Cancer as a robust system: implications for anticancer therapy. *Nat Rev Cancer, 4*(3), 227-235.

Koepfli, K. P., Paten, B., Genome, K. C. o. S., & O'Brien, S. J. (2015). The Genome 10K Project: a way forward. *Annu Rev Anim Biosci, 3*, 57-111. doi:10.1146/annurev-animal-090414-014900

Koes, D. R., & Camacho, C. J. (2012). PocketQuery: protein-protein interaction inhibitor starting points from protein-protein interaction structure. *Nucleic Acids Res, 40*(Web Server issue), W387-392. doi:10.1093/nar/gks336

Kortemme, T., & Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A, 99*(22), 14116-14121. doi:10.1073/pnas.202485799

Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins, 65*(2), 392-406. doi:10.1002/prot.21117

Krishnanand, K. N., & Ghose, D. (2008). Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. *Swarm Intelligence, 3*(2), 87-124. doi:10.1007/s11721-008-0021-5

Kruger, D. M., Ignacio Garzon, J., Chacon, P., & Gohlke, H. (2014). DrugScorePPI knowledge-based potentials used as scoring and objective function in protein-protein docking. *PLoS One, 9*(2), e89466. doi:10.1371/journal.pone.0089466

Kryshtafovych, A., Fidelis, K., & Moult, J. (2014). CASP10 results compared to those of previous CASP experiments. *Proteins, 82 Suppl 2*, 164-174. doi:10.1002/prot.24448

Kuhner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., . . . Gavin, A. C. (2009). Proteome organization in a genome-reduced bacterium. *Science, 326*(5957), 1235-1240. doi:10.1126/science.1176343

Kundrotas, P. J., Zhu, Z., Janin, J., & Vakser, I. A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proc Natl Acad Sci U S A, 109*(24), 9438-9441. doi:10.1073/pnas.1200678109

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J Mol Biol, 161*(2), 269-288. doi:http://dx.doi.org/10.1016/0022-2836(82)90153-X

Kuser, P., Cupri, F., Bleicher, L., & Polikarpov, I. (2008). Crystal structure of yeast hexokinase PI in complex with glucose: A classical "induced fit" example revised. *Proteins, 72*(2), 731-740. doi:10.1002/prot.21956

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature, 470*(7333), 187-197.

Laurie, A. T., & Jackson, R. M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics, 21*(9), 1908-1916. doi:10.1093/bioinformatics/bti315

Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics, 10*, 168. doi:10.1186/1471-2105-10-168

Lemmon, M. A., & Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. *Cell, 141*(7), 1117-1134. doi:10.1016/j.cell.2010.06.011

Lensink, M. F., & Wodak, S. J. (2014). Score_set: a CAPRI benchmark for scoring protein complexes. *Proteins, 82*(11), 3163-3169. doi:10.1002/prot.24678

Levy, E. D. (2010). A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol, 403*(4), 660-670. doi:10.1016/j.jmb.2010.09.028

Levy, E. D., Pereira-Leal, J. B., Chothia, C., & Teichmann, S. A. (2006). 3D complex: a structural classification of protein complexes. *PLoS Comput Biol, 2*(11), e155. doi:10.1371/journal.pcbi.0020155

Li, G., Li, M., Zhang, Y., Wang, D., Li, R., Guimera, R., . . . Zhang, M. Q. (2014). ModuleRole: a tool for modulization, role determination and visualization in protein-protein interaction networks. *PLoS One, 9*(5), e94608. doi:10.1371/journal.pone.0094608

Li, L., Chen, R., & Weng, Z. (2003). RDOCK: refinement of rigid-body protein docking predictions. *Proteins, 53*(3), 693-707. doi:10.1002/prot.10460

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., . . . Vidal, M. (2004). A map of the interactome network of the metazoan C. elegans. *Science, 303*(5657), 540-543. doi:10.1126/science.1091403

Lilley, D. M. J., & Eckstein, F. (2007). Chapter 1 Ribozymes and RNA Catalysis: Introduction and Primer *Ribozymes and RNA Catalysis* (pp. 1-10): The Royal Society of Chemistry.

Lise, S., Buchan, D., Pontil, M., & Jones, D. T. (2011). Predictions of hot spot residues at protein-protein interfaces using support vector machines. *PLoS One, 6*(2), e16774. doi:10.1371/journal.pone.0016774

Liu, S., Gao, Y., & Vakser, I. A. (2008). DOCKGROUND protein-protein docking decoy set. *Bioinformatics, 24*(22), 2634-2635. doi:10.1093/bioinformatics/btn497

Lo, Y. S., Chen, Y. C., & Yang, J. M. (2010). 3D-interologs: an evolution database of physical protein- protein interactions across multiple genomes. *BMC Genomics, 11 Suppl 3*, S7. doi:10.1186/1471-2164-11-S3-S7

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics, 19*(10), 1275-1283.

Luciani, D., & Bazzoni, G. (2012). From networks of protein interactions to networks of functional dependencies. *BMC Syst Biol, 6*, 44. doi:10.1186/1752-0509-6-44

Ma'ayan, A. (2009). Insights into the organization of biochemical regulatory networks using graph theory analyses. *The Journal of biological chemistry, 284*(9), 5451-5455. doi:10.1074/jbc.R800056200

MacBeath, G. (2002). Protein microarrays and proteomics. *Nat Genet, 32 Suppl*, 526-532. doi:10.1038/ng1037

Maciag, K., Altschuler, S. J., Slack, M. D., Krogan, N. J., Emili, A., Greenblatt, J. F., . . . Wu, L. F. (2006). Systems-level analyses identify extensive coupling among gene expression machines. *Mol Syst Biol, 2*, 2006 0003. doi:10.1038/msb4100045

Marion, D. (2013). An introduction to biological NMR spectroscopy. *Mol Cell Proteomics, 12*(11), 3006-3025. doi:10.1074/mcp.O113.030239

Mashiach, E., Nussinov, R., & Wolfson, H. J. (2010). FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins, 78*(6), 1503-1519. doi:10.1002/prot.22668

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., . . . Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res, 11*(12), 2120-2126. doi:10.1101/gr.205301

Moal, I. H., & Bates, P. A. (2010). SwarmDock and the use of normal modes in protein-protein docking. *Int J Mol Sci, 11*(10), 3623-3648. doi:10.3390/ijms11103623

Moal, I. H., & Fernandez-Recio, J. (2012). SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics, 28*(20), 2600-2607. doi:10.1093/bioinformatics/bts489

Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2007). Computational alanine scanning mutagenesis--an improved methodological approach. *J Comput Chem, 28*(3), 644-654. doi:10.1002/jcc.20566

Morrison, K. L., & Weiss, G. A. (2001). Combinatorial alanine-scanning. *Curr Opin Chem Biol, 5*(3), 302-307.

Mosca, R., Ceol, A., & Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nat Methods, 10*(1), 47-53. doi:10.1038/nmeth.2289

Mosca, R., Ceol, A., Stein, A., Olivella, R., & Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res, 42*(Database issue), D374-379. doi:10.1093/nar/gkt887

Mosca, R., Pons, C., Fernandez-Recio, J., & Aloy, P. (2009). Pushing structural information into the yeast interactome by high-throughput protein docking experiments. *PLoS Comput Biol, 5*(8), e1000490. doi:10.1371/journal.pcbi.1000490

Mosca, R., Tenorio-Laranga, J., Olivella, R., Alcalde, V., Ceol, A., Soler-Lopez, M., & Aloy, P. (2015). dSysMap: exploring the edgetic role of disease mutations. *Nat Methods, 12*(3), 167-168. doi:10.1038/nmeth.3289

Muller, A., MacCallum, R. M., & Sternberg, M. J. (2002). Structural characterization of the human proteome. *Genome Res, 12*(11), 1625-1641. doi:10.1101/gr.221202

Negroni, J., Mosca, R., & Aloy, P. (2014). Assessing the applicability of template-based protein docking in the twilight zone. *Structure, 22*(9), 1356-1362. doi:10.1016/j.str.2014.07.009

Niroula, A., Urolagin, S., & Vihinen, M. (2015). PON-P2: prediction method for fast and reliable identification of harmful variants. *PLoS One, 10*(2), e0117380. doi:10.1371/journal.pone.0117380

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report, 2*(3), 117-120. doi:10.1089/152791601750294344

Ofran, Y., & Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol, 3*(7), e119. doi:10.1371/journal.pcbi.0030119

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res, 27*(1), 29-34. doi:10.1093/nar/27.1.29

Pavelka, A., Chovancova, E., & Damborsky, J. (2009). HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res, 37*(Web Server issue), W376-383. doi:10.1093/nar/gkp410

Pierce, B., & Weng, Z. (2007). ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins, 67*(4), 1078-1086. doi:10.1002/prot.21373

Pizzuti, C., & Rombo, S. E. (2014). Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics, 30*(10), 1343-1352. doi:10.1093/bioinformatics/btu034

Pons, C., Talavera, D., de la Cruz, X., Orozco, M., & Fernandez-Recio, J. (2011). Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. *J Chem Inf Model, 51*(2), 370-377. doi:10.1021/ci100353e

Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., . . . Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics, 29*(7), 845-854. doi:10.1093/bioinformatics/btt055

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., . . . Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods, 24*(3), 218-229. doi:10.1006/meth.2001.1183

Qiu, Y.-Q. (2013). KEGG Pathway Database. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 1068-1069). New York, NY: Springer New York.

Redington, P. K. (1992). MOLFIT: A computer program for molecular superposition. *Computers & Chemistry, 16*(3), 217-222. doi:10.1016/0097-8485(92)80005-k

Riera, C., Padilla, N., & de la Cruz, X. (2016). The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. *Hum Mutat, 37*(10), 1013-1024. doi:10.1002/humu.23048

Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., . . . Roden, D. M. (2010). Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *The American Journal of Human Genetics, 86*(4), 560-572. doi:10.1016/j.ajhg.2010.03.003

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., . . . Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell, 159*(5), 1212-1226. doi:10.1016/j.cell.2014.10.050

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., . . . Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature, 437*(7062), 1173-1178. doi:10.1038/nature04209

Russell, R. J., Haire, L. F., Stevens, D. J., Collins, P. J., Lin, Y. P., Blackburn, G. M., . . . Skehel, J. J. (2006). The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature, 443*(7107), 45-49. doi:10.1038/nature05114

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., . . . Vidal, M. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell, 161*(3), 647-660. doi:10.1016/j.cell.2015.04.013

Salomon-Ferrer, R., Case, D. A., & Walker, R. C. (2013). An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 3*(2), 198-210. doi:10.1002/wcms.1121

Sant'Anna, R., Gallego, P., Robinson, L. Z., Pereira-Henriques, A., Ferreira, N., Pinheiro, F., . . . Ventura, S. (2016). Repositioning tolcapone as a potent inhibitor of transthyretin amyloidogenesis and associated cellular toxicity. *Nat Commun, 7*, 10787. doi:10.1038/ncomms10787

Schiffer, M., Ainsworth, C., Xu, Z. B., Carperos, W., Olsen, K., Solomon, A., . . . Chang, C. H. (1989). Structure of a second crystal form of Bence-Jones protein Loc: strikingly different domain associations in two crystal forms of a single protein. *Biochemistry, 28*(9), 4066-4072.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., & Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res, 33*(Web Server issue), W363-367. doi:10.1093/nar/gki481

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res, 33*(Web Server issue), W382-388. doi:10.1093/nar/gki387

Scott, A. F., Amberger, J., Brylawski, B., & McKusick, V. A. (1999). OMIM: Online Mendelian Inheritance in Man. In S. Letovsky (Ed.), *Bioinformatics: Databases and Systems* (pp. 77-84). Boston, MA: Springer US.

Segura, J., & Fernandez-Fuentes, N. (2011). PCRPi-DB: a database of computationally annotated hot spots in protein interfaces. *Nucleic Acids Res, 39*(Database issue), D755-760. doi:10.1093/nar/gkq1068

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., . . . Rubio, A. (2005). Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform, 2*(4), 330-338. doi:10.1109/TCBB.2005.50

Shah, N., Sattar, A., Benanti, M., Hollander, S., & Cheuck, L. (2006). Magnetic Resonance Spectroscopy as an Imaging Tool for Cancer: A Review of the Literature. *The Journal of the American Osteopathic Association, 106*(1), 23-27. doi:10.7556/jaoa.2006.106.1.23

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., . . . Aparicio, S. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature, 486*(7403), 395-399. doi:10.1038/nature10933

Shingate, P., Manoharan, M., Sukhwal, A., & Sowdhamini, R. (2014). ECMIS: computational approach for the identification of hotspots at protein-protein interfaces. *BMC Bioinformatics, 15*, 303. Retrieved from doi:10.1186/1471-2105-15-303

Shoemaker, B. A., Zhang, D., Tyagi, M., Thangudu, R. R., Fong, J. H., Marchler-Bauer, A., . . . Panchenko, A. R. (2012). IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res, 40*(Database issue), D834-840. doi:10.1093/nar/gkr997

Shulman-Peleg, A., Shatsky, M., Nussinov, R., & Wolfson, H. J. (2007). Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol, 5*, 43. doi:10.1186/1741-7007-5-43

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res, 40*(Web Server issue), W452-457. doi:10.1093/nar/gks539

Sinha, R., Kundrotas, P. J., & Vakser, I. A. (2012). Protein docking by the interface structure similarity: how much structure is needed? *PLoS One, 7*(2), e31349. doi:10.1371/journal.pone.0031349

Smyth, M. S., & Martin, J. H. (2000). x ray crystallography. *Molecular pathology : MP, 53*(1), 8-14. doi:10.1136/mp.53.1.8

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., . . . Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell, 122*(6), 957-968. doi:10.1016/j.cell.2005.08.029

Stingele, S., Stoehr, G., Peplowska, K., Cox, J., Mann, M., & Storchova, Z. (2012). Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol Syst Biol, 8*(1), 608. doi:10.1038/msb.2012.40

Strogatz, S. H. (2001). Exploring complex networks. *Nature, 410*(6825), 268-276. doi:10.1038/35065725

Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., & Wiuf, C. (2008). Estimating the size of the human interactome. *Proc Natl Acad Sci U S A, 105*(19), 6959-6964. doi:10.1073/pnas.0708078105

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., . . . von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res, 43*(Database issue), D447-452. doi:10.1093/nar/gku1003

Teichmann, S. A. (2002). Principles of protein-protein interactions. *Bioinformatics, 18 Suppl 2*(suppl_2), S249. doi:10.1093/bioinformatics/18.suppl_2.S249

Teng, S., Madej, T., Panchenko, A., & Alexov, E. (2009). Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. *Biophys J, 96*(6), 2178-2188. doi:10.1016/j.bpj.2008.12.3904

The International HapMap, C. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi:http://www.nature.com/nature/journal/v437/n7063/suppinfo/nature04226_S1.html

Tobi, D. (2010). Designing coarse grained-and atom based-potentials for protein-protein docking. *BMC Struct Biol, 10*(1), 40. doi:10.1186/1472-6807-10-40

Tobi, D., & Bahar, I. (2006). Optimal design of protein docking potentials: efficiency and limitations. *Proteins, 62*(4), 970-981. doi:10.1002/prot.20859

Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., . . . Feolo, M. (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res, 42*(Database issue), D975-979. doi:10.1093/nar/gkt1211

Tuncbag, N., Keskin, O., & Gursoy, A. (2010). HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res, 38*(Web Server issue), W402-406. doi:10.1093/nar/gkq323

Uetz, P., & Hughes, R. E. (2000). Systematic and large-scale two-hybrid screens. *Curr Opin Microbiol, 3*(3), 303-308. doi:https://doi.org/10.1016/S1369-5274(00)00094-1

UniProt, C. (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Res, 35*(Database issue), D193-197. doi:10.1093/nar/gkl929

Vajda, S., & Kozakov, D. (2009). Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol, 19*(2), 164-170. doi:10.1016/j.sbi.2009.02.008

Vavricka, C. J., Li, Q., Wu, Y., Qi, J., Wang, M., Liu, Y., . . . Gao, G. F. (2011). Structural and functional analysis of laninamivir and its octanoate prodrug reveals group specific

mechanisms for influenza NA inhibition. *PLoS Pathog, 7*(10), e1002249. doi:10.1371/journal.ppat.1002249

Vazquez, A., Flammini, A., Maritan, A., & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol, 21*(6), 697-700. doi:10.1038/nbt825

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., . . . Smith, H. O. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science, 304*(5667), 66-74. doi:10.1126/science.1093857

Vidal, M., Cusick, M. E., & Barabasi, A. L. (2011). Interactome networks and human disease. *Cell, 144*(6), 986-998. doi:10.1016/j.cell.2011.02.016

Viswanath, S., Ravikant, D. V., & Elber, R. (2013). Improving ranking of models for protein complexes with side chain modeling and atomic potentials. *Proteins, 81*(4), 592-606. doi:10.1002/prot.24214

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res, 31*(1), 258-261.

Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., . . . Weng, Z. (2015). Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol, 427*(19), 3031-3041. doi:10.1016/j.jmb.2015.07.016

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., & Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotech, 30*(2), 159-164. doi:10.1038/nbt.2106

http://www.nature.com/nbt/journal/v30/n2/abs/nbt.2106.html#supplementary-information

Wells, J. A., & McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature, 450*(7172), 1001-1009. doi:http://www.nature.com/nature/journal/v450/n7172/suppinfo/nature06526_S1.html

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res, 42*(Database issue), D1001-1006. doi:10.1093/nar/gkt1229

Winter, C., Henschel, A., Kim, W. K., & Schroeder, M. (2006). SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res, 34*(Database issue), D310-314. doi:10.1093/nar/gkj099

Wu, X., & Chen, J. Y. (2012, 2-4 Dec. 2012). *An evaluation for merging signaling pathways by using protein-protein interaction data.* Paper presented at the Proceedings 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS).

Wu, X., Harrison, S. H., & Chen, J. Y. (2009). Pattern discovery in breast cancer specific protein interaction network. *Summit Transl Bioinform, 2009*, 1-5.

Wu, X., Hasan, M. A., & Chen, J. Y. (2014). Pathway and network analysis in proteomics. *J Theor Biol, 362*, 44-52. doi:10.1016/j.jtbi.2014.05.031

Wu, X., Huan, T., Pandey, R., Zhou, T., & Chen, J. Y. (2009). Finding fractal patterns in molecular interaction networks: a case study in Alzheimer's disease. *Int J Comput Biol Drug Des, 2*(4), 340-352. doi:10.1504/IJCBDD.2009.030765

Xia, J. F., Zhao, X. M., Song, J., & Huang, D. S. (2010). APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics, 11*, 174. doi:10.1186/1471-2105-11-174

Zanzoni, A., Soler-Lopez, M., & Aloy, P. (2009). A network medicine approach to human disease. *FEBS Lett, 583*(11), 1759-1765. doi:10.1016/j.febslet.2009.03.001

Zhang, C., & Lai, L. (2011). SDOCK: a global protein-protein docking program using stepwise force-field potentials. *J Comput Chem, 32*(12), 2598-2612. doi:10.1002/jcc.21839

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics, 9*(1), 40. doi:10.1186/1471-2105-9-40

Zhang, Y. (2014). Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins, 82 Suppl 2*, 175-187. doi:10.1002/prot.24341

Zhong, Q., Simonis, N., Li, Q. R., Charloteaux, B., Heuze, F., Klitgord, N., . . . Vidal, M. (2009). Edgetic perturbation models of human inherited disorders. *Mol Syst Biol, 5*(1), 321. doi:10.1038/msb.2009.80

Zhu, X., & Mitchell, J. C. (2011). KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins, 79*(9), 2671-2683. doi:10.1002/prot.23094

FIGURE LEGENDS

Figure 1. Study of genetic variants: reductionist vs. systemic approach. A genetic variant like a nsSNP can modify the structure and/or dynamics of a protein at molecular level, which in turn can alter the interaction network of such protein. The impact in phenotype can be studied just at molecular level (reductionist approach) but very often the phenotypic observation is more linked with the effects at higher levels of organization, such as interaction network or cell scales (systemic approach).

Figure 2. Interaction network of the proteins involved in the RAS-MAPK cascade. In red, proteins hosting mutations associated with cancer or RASopathies. They are part of a larger interaction network (generated with Interactome3D; http://interactome3d.irbbarcelona.org/), involving many more proteins represented here as blue circles.

Figure 3. Basic scheme of a protein-protein docking method. From the coordinates of two interacting proteins, computational docking generates thousands of possible complex models, ideally containing near-native models. A scoring scheme based on energetic terms or empirical potentials will try to identify such correct models.

Figure 4. Comparison of *ab initio* and template-based docking approaches. *Ab initio* docking aims to build a protein-protein complex from the structures of the individual interacting proteins, using computational sampling and scoring based on energy considerations or empirical

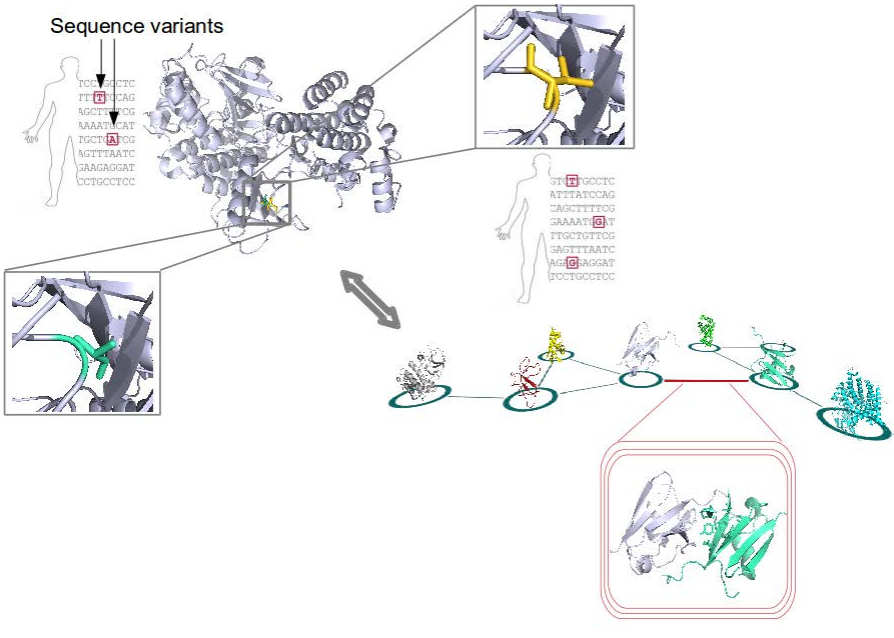
parameters. In template-based docking, the protein-protein complex structure is built based on a template complex structure in which the components are homologous to the individual interacting proteins.

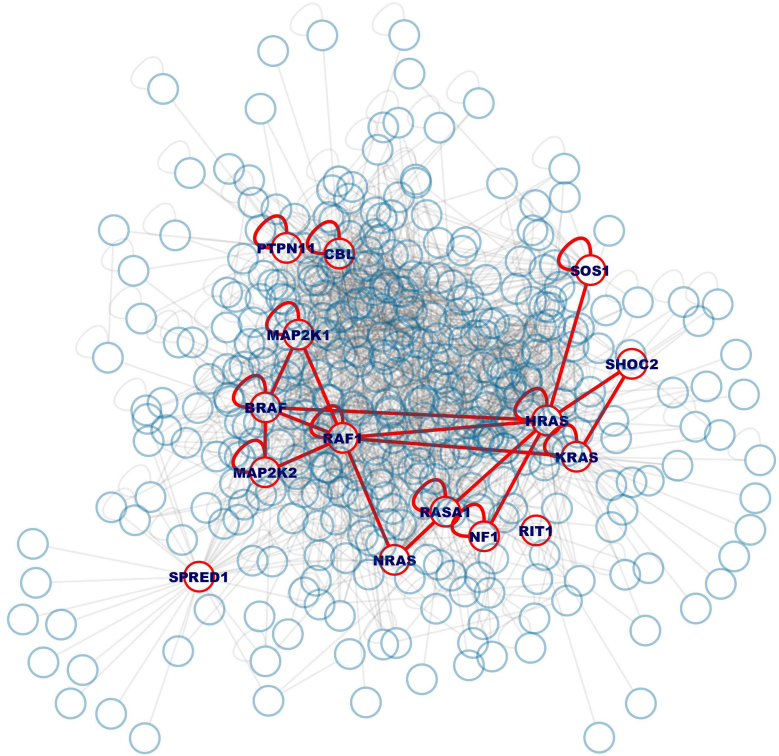
Figure 5. Different types of zones in a protein-protein interface. In a protein-protein complex structure, the interface residues can be defined as those that become more buried upon binding. Such interface residues can be classified in: i) core, residues exposed in the unbound state and buried in the complex; ii) rim, residues exposed in the unbound state and slightly less exposed in the complex; and iii) support, residues buried in the unbound state and more buried in the complex.

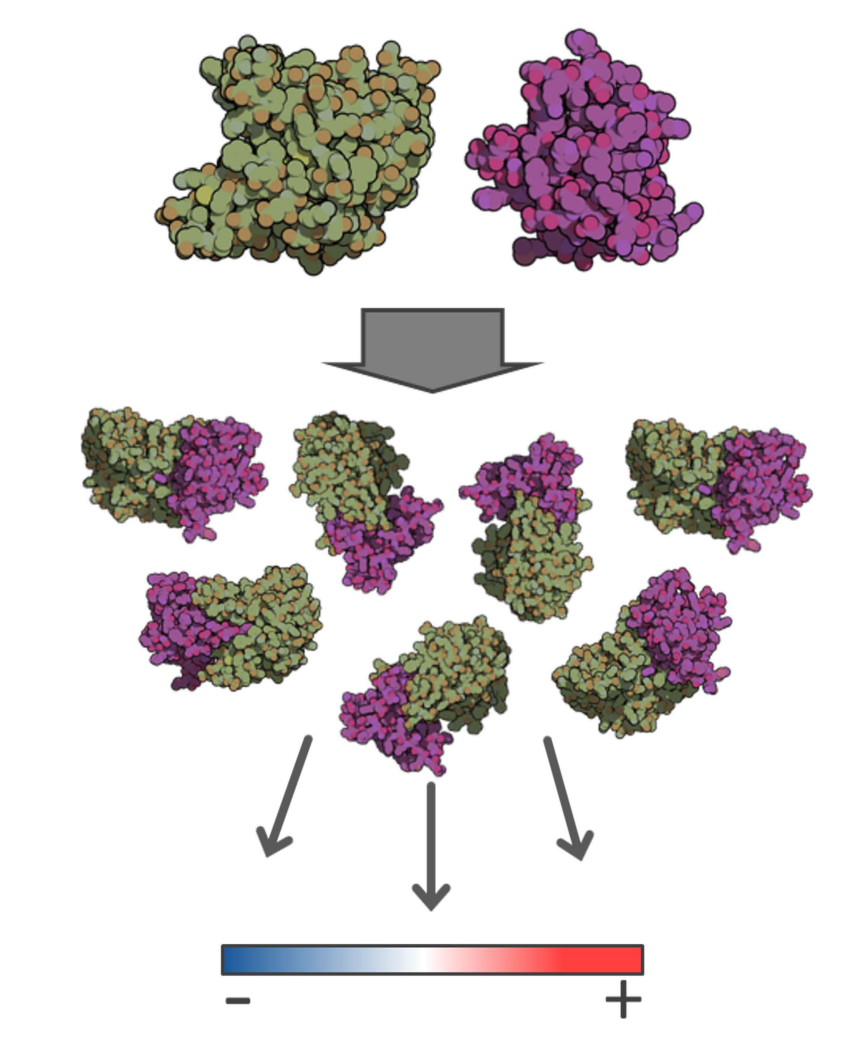
Figure 6. New advances in the interpretation of biological mutational data: from molecules to networks.

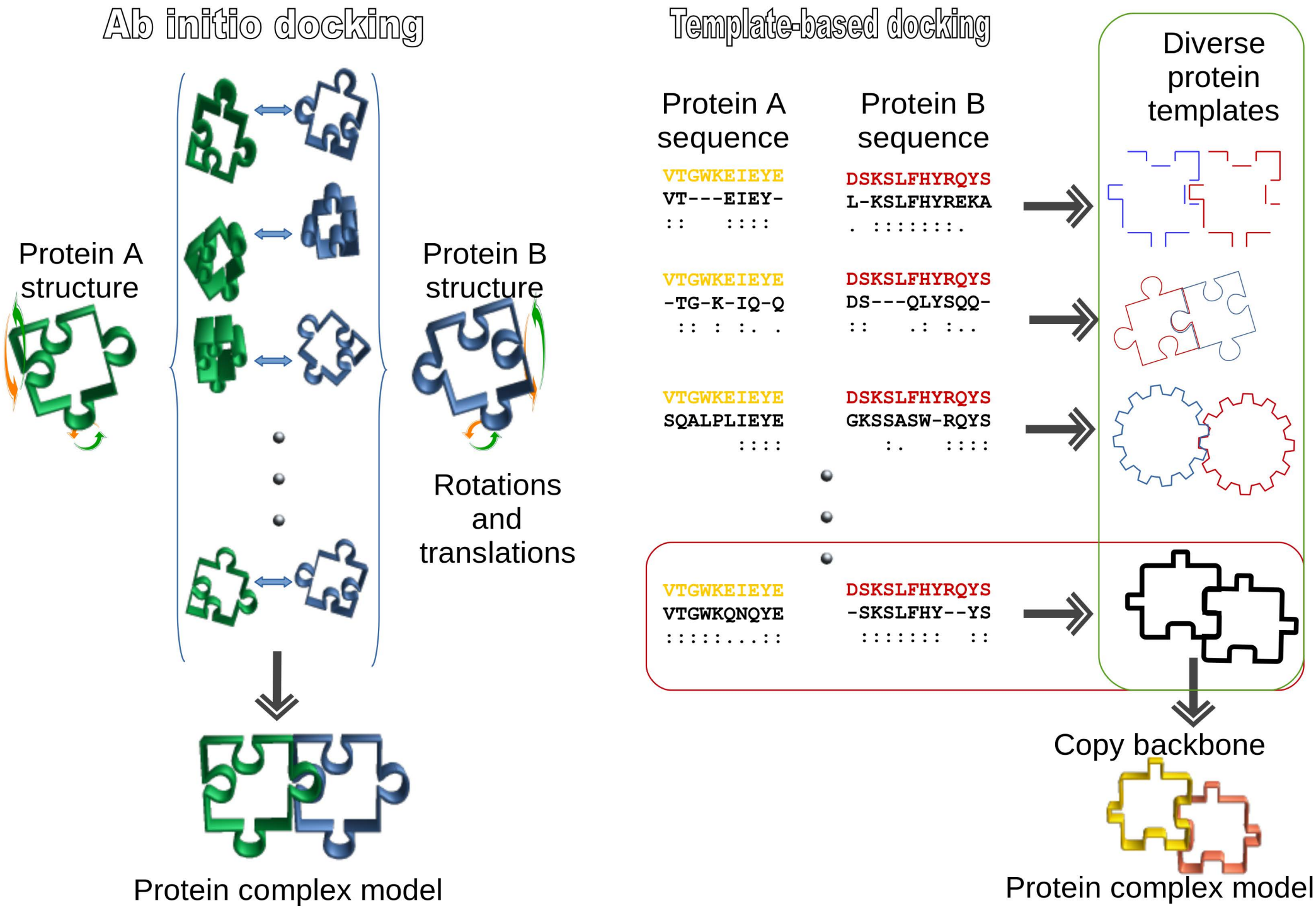
Personalized medicine aims to understand the effect of genetic variants on the development and onset of pathological conditions and on the response to existing treatments, so that therapeutic intervention can be optimally selected for each patient given his genetic profile. In this context, disease needs to be described at all levels, from molecular to system, for which understanding the interaction networks involved in a given pathology is a key step. Computational methods like protein-protein docking are useful tools for characterizing protein-protein interfaces to localize pathological variants and predict its effects on binding affinity (left panel). The effect of this molecular perturbation on the interaction networks (edgetic effect) can lead to the development of more accurate personalized therapies and the developments of new drugs (right panel).

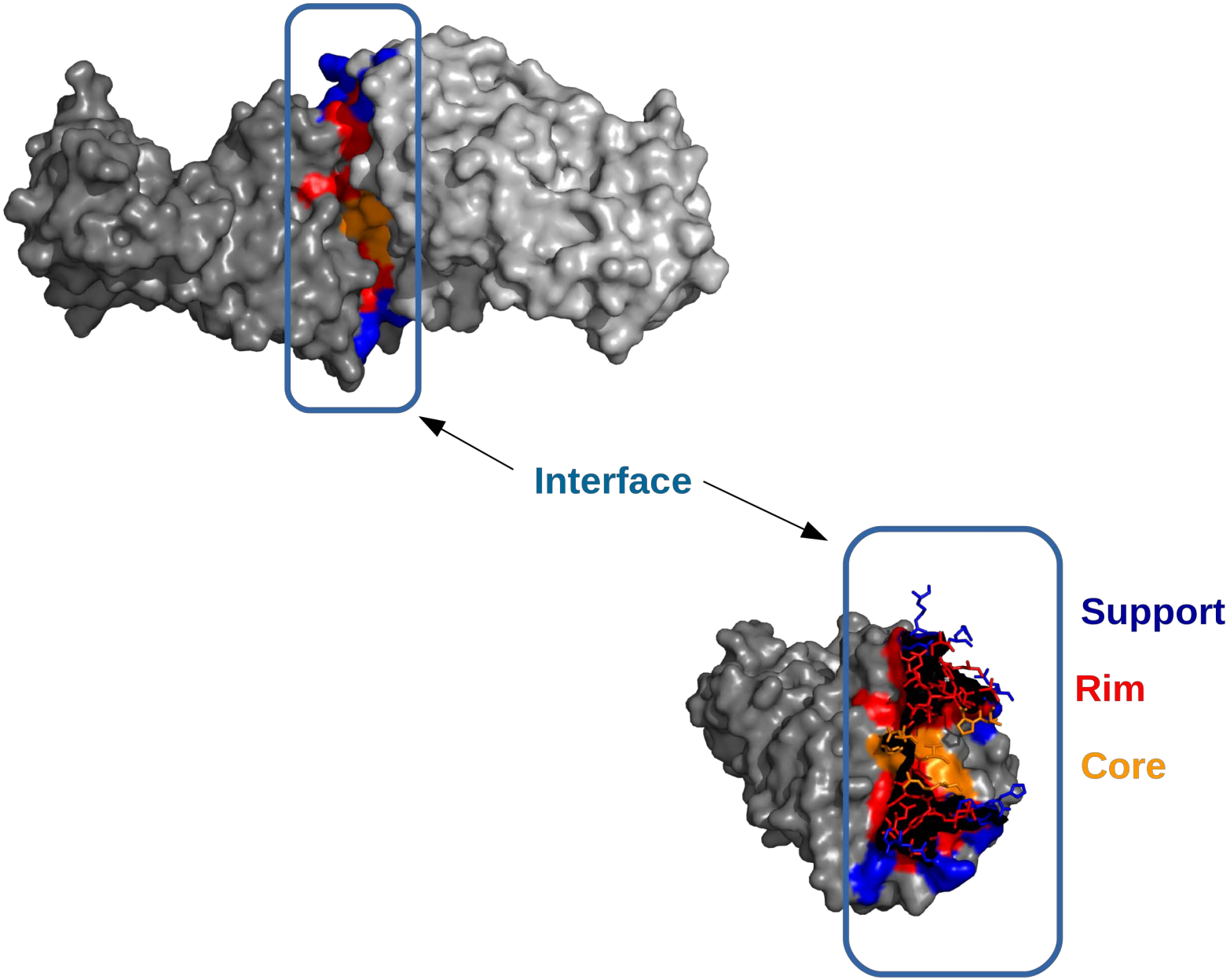
Figure 7. Inhibiting protein-protein interactions with small molecules. Two examples of protein-protein interactions that are known to be inhibited by small molecules: IL2R/IL2 complex (left) and XIAP-Bir3/CASPASE9 complex (right). The crystallographic structures of the proteins bound to the corresponding small-molecule inhibitors are shown (PDB codes 1PY2 and 1TFT, respectively), with details of the protein-inhibitor interface (cyan surface). For comparison, the orientation of the partner protein in the corresponding protein-protein complex structure is shown (orange ribbon; PDB codes 1Z92 and 1NW9, respectively). In both cases, the small molecule clearly overlaps with part of the protein-protein interface (blue surface), and therefore disrupt the interaction. Inhibitor cavities are not fully open in the unbound protein (bottom panel; PDB codes 1M47 and 1F9X, respectively), which makes them highly difficult to identify.


Sequence variants

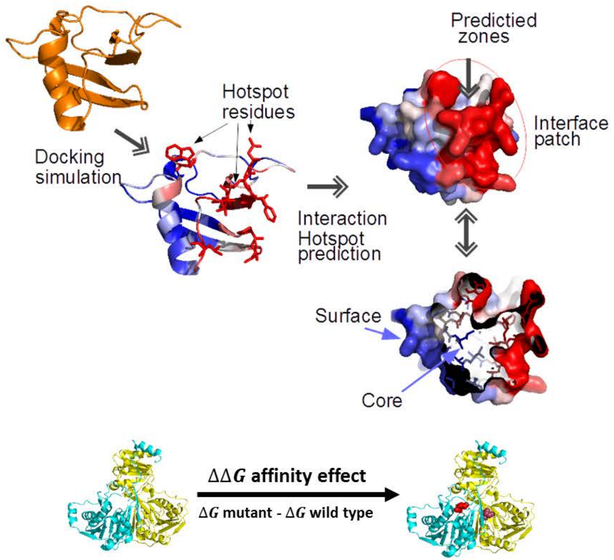

PTPN11
CBL
SOS1
MAP2K1
SHOC2
BRAF
HRAS
RAF1
KRAS
MAP2K2
RASA1
NF1
RIT1
NRAS
SPRED1

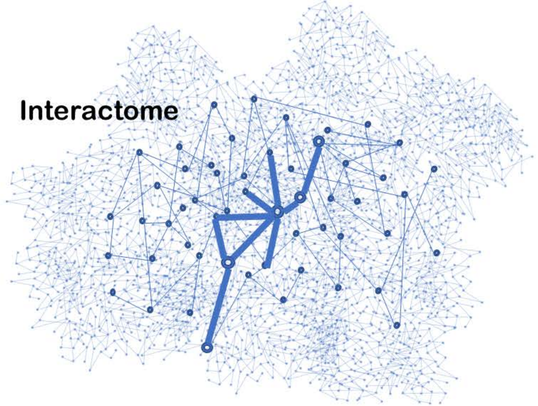


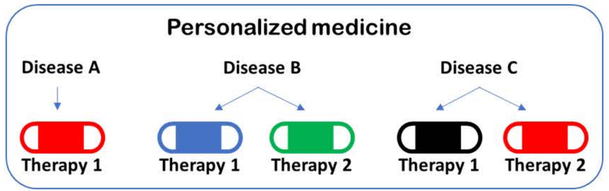

Ab initio docking
Protein A structure
Protein B structure
Rotations and translations
Protein complex model
Template-based docking
Protein A sequence
Protein B sequence
Diverse protein templates
VTGWKEIEYE
VT---EIEY-
DSKSLFHYRQYS
L-KSLFHYREKA
VTGWKEIEYE
-TG-K-IQ-Q
DSKSLFHYRQYS
DS---QLYSQQ-
VTGWKEIEYE
SQALPLIEYE
DSKSLFHYRQYS
GKSSASW-RQYS
VTGWKEIEYE
VTGWKQNQYE
DSKSLFHYRQYS
-SKSLFHY--YS
Copy backbone
Protein complex model

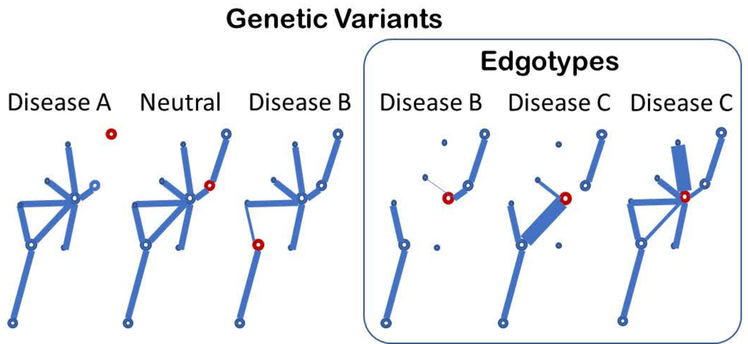

Interface
Support
Rim
Core

Characterizing protein-protein interfaces by docking









Interpretation of pathological mutations at interactomic level