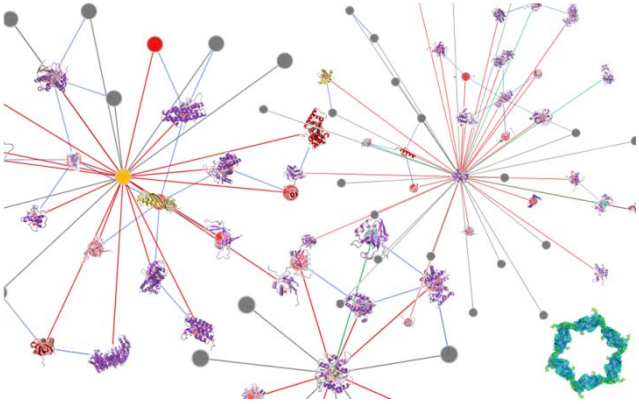
Università degli Studi di Roma “La Sapienza”

“Detecting mutually exclusive interactions in protein-protein interaction maps”.



Tutore
Prof.ssa Anna Tramontano

Docente guida
Prof.ssa Anna Tramontano

Coordinatore
Prof. Marco Tripodi

Dottoranda
Carmen Sánchez Claros

Dottorato di Ricerca in Scienze Pasteuriane
XXIV CICLO

A mis padres

Acknowledgements

It is difficult to summarize in a few lines my sincere gratitude towards people who have helped me during these three years. Without them it would have been impossible to accomplish this thesis, in which I have put so much hope and efforts.

I am especially thankful to Anna Tramonatano for giving me the opportunity to join the Biocomputing group at the Sapienza University of Rome, and for her dedicate supervision and assistance. I would also like to thank the Pasteurian Sciences PhD School for useful seminars and discussions and the King Abdullah University of Science and Technology for the funding.

Por último, en el apartado personal, mi más sincera gratitud a toda mi familia, en especial a mis padres, por su amor, comprensión y dedicación constantes que permanentemente me hacen y me han hecho sentir a pesar de la distancia. Os quiero muchísimo.

No puedo olvidar a mi compañero y amigo Ahmed Sayadi con el que he compartido laboratorio e incontables horas en la “mensa”. Gracias por los buenos y malos momentos, por aguantarme y por escucharme.

Pero especialmente quisiera darle las gracias a una persona que siempre ha confiado en mí, que ha estado siempre a mi lado en tantísimos momentos duros a lo largo de estos tres años, a mi compañero de aventuras, Daniel Carbajo, sin él nada de esto hubiese sido posible. Nunca olvidaré, nuestros interminables paseos por Roma y nuestro pequeño oasis, “Villa Dominici”.

Arrivederci Roma.

Index

Introduction

Protein-Protein Interactions

Proteins are responsible for an impressively large variety of functions, they are involved in catalytic reactions, transportation of ions, molecules and macromolecules across the membranes, structural components of the cell, traverse the membranes to yield regulated channels, and transmit the information from the DNA to the RNA. They synthesize new molecules, and are responsible for their degradation. Proteins are the vehicles of the immune response and of viral entry into cells (Keskin, et al., 2008). All these functions are realized through interactions with many other molecules, small molecules, DNA, RNA and proteins.

To properly understand the significance of protein-protein interactions in the cell it is important to addresses two problems: first, the identification of the different interactions that involve each biological function, and second to determine how the involved proteins interact and which are the consequences of the interaction.

Protein-protein interactions can be classified as physical or logical. Physical protein interactions are those interactions that happen when two proteins form a complex. Physical interactions are, for example, stable complexes where the functional unit is formed by the assembly of more than one protein; this is any protein in the ribosome or in the trascriptome machinery that is in contact with other proteins in the complex. However, not all the proteins interact physically; they can affect indirectly other proteins (logical interactions) by regulating their expression, or by changing the concentration of a factor that, in turn, is sensed by the target protein. The two modes of interaction are not exclusive. The same proteins can interact both physically and logically. However, the existence of logical interactions between proteins should not be confused with physical interactions.

Protein-protein physical interactions are structurally and functionally diverse. Nooren and Thornton in 2003 published the first classification of protein complexes with respect to three different properties: composition, stability of the components and duration of the association. With respect to the composition of the complex, protein-protein interactions can occur between identical or non-identical subunits or chains. If the interaction happens between identical chains, the complex is classified

as a homo-oligomer. If the composition of the complex includes chains that are different it is called a hetero-oligomer. We can further divide homo-oligomers with respect to their structural symmetry (Goodsell & Olson, 2000). Isologous associations comprise the same interacting surface on both monomers. In contrast, heterologous associations involve different interacting interfaces that, in some cases, can lead to infinite aggregation (Monod, 1965).

In terms of stability of the subunits two different types of complexes can be distinguished. Obligate complexes are formed by chains or subunits that are unstable on their own and cannot exist independently i.e. obligatory complexes only function when associated in the complex, while the components of non-obligate complexes are stable on their own, and association is not required for stability. Another possible classification has to do with the duration of a complex; there are complexes that are transient and others that are essentially permanent.

Methods for Detection and Analysis of Protein-Protein Interactions

The broad recognition of the importance of characterizing all protein interactions in a cell has been extensively studied in various scientific disciplines, such as biochemistry, genomics, bioinformatics, computational molecular modeling, cellular and molecular biology, biophysics, etc. These approaches can be divided into experimental and computational approaches.

Experimental approaches

Experimental methods can be divided into two different categories: methods that are designed to identify and validate a small number of interactions, and methods that involve the screening of large scale protein interactions i.e. high-throughput experiments.

Methods that identify small number of target interactions include X-ray crystallography, NMR spectroscopy, fluorescence resonance energy transfer (Yan & Marriott, 2003) and surface plasmon resonance (Karlsson, 2004). These techniques,

which are so-called biophysical approaches, are the same that "wet" laboratories use to determine the structure of proteins. They are time and labor consuming and not all structures of proteins or protein complexes can be experimentally determined.

High-throughput experiments are based on a common principle which is similar to the principle of fishing, hence the terminology: bait protein, prey protein, and molecular fishing. A bait protein is a known protein used by the experimenter to "catch" and identify one or several unknown protein-protein interaction partner proteins which are called prey proteins (Ivanov, et al., 2010). Experimental high-throughput methods can be classified in Genomic and Biochemical Approaches (Berggård, et al., 2007).

Genomic Approaches are sophisticated strategies that are designed to discover genes that show interactions with other genes, which can encode proteins that physically interact with proteins encoded by the known genes. In other cases, genetic methods can be used to confirm interactions among previously identified proteins. These strategies are the yeast two-hybrid system, the synthetic genetic arrays and the synexpression.

The yeast two-hybrid system (Chien, et al., 1991; Fields & Song, 1989; Fields & Sternglanz, 1994) is a genetic method that uses transcriptional activity as a measure of protein-protein interaction. It relies on the modular nature of many site-specific transcriptional activators that have two domains (Brent & Ptashne, 1985; Hope & Struhl, 1986; Keegan, et al., 1986): a DNA binding domain and a transcription activation domain. This approach requires that two hybrids are constructed: a bait protein which is a fusion between the target protein X and the DNA binding domain of a transcription factor (DBD), and the prey protein(s) Y fused together with the transcription activation domain of the same transcription factor (AD). These two constructs are expressed in a cell containing a reporter gene. If the DBD-X fusion protein binds to the operator site in the promoter region it cannot alone activate the transcription of the reporter gene because of the absence of the AD domain. Hence, the DBD-X fusion protein must interact with its binding partner AD-Y, to form a complete transcription factor that allows the induction of the reporter gene. There are a variety of versions of this approach, comprising always a transcription factor, Gal4 protein in *Saccharomyces cerevisiae* or LexA protein in *Escherichia coli*, and reporter genes, usually lacZ coding for β-galactosidase that can be easily detected.

As a genetic system, the yeast two-hybrid system is well suited to high-throughput applications that require automatization. Large-scale two hybrid approaches have used two complementary approaches, the matrix approach and the library screening approach for screening large sets of proteins. In the matrix approach, a yeast strain expressing the bait protein of interest is associated with an array of yeast strains that express many different prey proteins. The interactions between proteins can be detected by growing the mated strains on a selective medium in an array, which allows the identification of the growing colonies where the prey protein interacts with the bait protein of interest. In the publication by Uetz et al. (Uetz, et al., 2000), a yeast two-hybrid matrix experiment was performed by merging a pool of 6,000 yeast transformants, with each transformant expressing one of the open reading frames as a fusion to an activation domain ,with cells transformed with one given BD plasmid. In such manner, they have identified 281 interactions. The library screening approach does not separate the different prey strains on an array but instead screens a set of baits against a library. For example, Fromont-Racine, et al., in 1997 generated a highly complex library of random yeast genomic fragments containing approximately 3000000 full-length open reading frames and fragments. The idea of including fragments is because some interactions take place between single domains. The fragments, as well as the full-length open reading frames, are cloned into an AD vector. The resulting transformants are then

collected into aliquots, each of which constitute representatives of the complete original library. In the same work previously mentioned for matrix approaches, Uetz et al. (Uetz, et al., 2000) performed a library screen using their set of baits and preys. Out of a total of more than 5341 open reading frames tested, 817 were identified as putative protein-protein interactions, thus identifying a grand total of 692 interacting pairs. Besides, High-throughput yeast two-hybrid screens have been carried out for *Plasmodium falciparum* (LaCount, et al., 2005), *Caenorhabditis elegans* (Li, et al., 2004), *Drosophila melanogaster* (Giot, et al., 2003), and more recently, approximately a third of the Homo sapiens genome has been screened in this manner (Rual, et al., 2005; Stelzl, et al., 2005).

The Synthetic Genetic Arrays (SGA), as well as the two-hybrid systems, belong to the group of genomic approaches and aim at a large scale analysis of genetic relationships by systematic construction of double mutants (Tong, et al., 2001; Tong, et al., 2004; Tong & Boone, 2006). This approach is based on the observation (Tong, et al., 2001) that more than 80% of the approximately 6200 genes of the yeast *Saccharomyces cerevisiae* are not essential. However, even non-essential genes can cause lethality when two of them are mutated at the same time, forming a synthetic lethal interaction. These genes might encode proteins that interact physically. The SGA analysis offers an efficient approach for the systematic

construction of double mutants and enables a global analysis of synthetic lethal genetic interactions. A typical SGA analysis consists of a matrix of combinations of the mutation in the target gene and more than 5000 mutations in viable strains with single-gene deletions aimed at the detection of double mutants with a defective phenotype. SGA is an in-vivo approach producing large amount of protein interaction data, it is useful to perform unbiased genome-wide screens.

Correlating mRNA expression profiles or synexpression is based on the analysis of correlations between transcriptomic or interactomic data (Ge, et al., 2001). Transcriptomic data, mRNA levels, are regularly measured under a variety of different cellular conditions, and genes are clustered if they show a similar transcriptional response to these conditions (clusters of gene expression profile). These clusters might encode physically interacting proteins, interactomic data. The example of *Saccharomyces cerevisiae* shows that genes with similar expression profiles often code for interacting proteins. The synexpression is an in vivo technique that allows a large coverage of different cellular conditions and is a powerful method to discriminate among cell states or disease outcomes.

Biochemical Approaches for identifying interacting proteins are varied and time-honored, some being as old as the field of protein chemistry itself. Biochemical approaches for the isolation and characterization of protein–protein interactions includes Coinmunoprecipitation, surface plasmon resonance and mass spectrometric analysis.

Column chromatography is a method used to purify individual proteins from complex mixtures, passing these through a column containing a porous solid matrix. The different proteins are retained by their interaction with the matrix, and they can be collected separately. Depending on the choice of matrix, proteins can be separated according to their charge (ion-exchange chromatography), their hydrophobicity (hydrophobic chromatography), their size (gel-filtration chromatography), or their ability to bind to particular small molecules or to other macromolecules (affinity chromatography). The last one, affinity chromatography, takes advantage of the biologically important binding interactions that occur on protein surfaces. If a substrate molecule (for example, a specific ligand or an antibody molecule) is covalently linked to an inert insoluble matrix, the protein that binds this specific substrate is specifically retained by the matrix and can next be eluted in nearly pure form. There are different classes of affinity targets, as well as different purification goals.

In the Coinmunoprecipitation technique, specific antibodies are used for the isolation of bait proteins bound to partner proteins from cell lysates (Phizicky & Fields, 1995; Masters, 2004; Yaciuk, 2007; Free, et al., 2009). The choice of antibodies is a key step in this technique, because they have a high affinity to target proteins, (in order to bind the antigen strongly enough and isolate it from the mixture) and high specificity (to minimize nonspecific interactions). Both monoclonal and polyclonal antibodies are used in these experiments. Other affinity reagents can be used for this, including chromatographic resins with conjugated glutathione for the isolation of partner proteins bound to a target protein tagged with glutathione-*S*-transferase (GST) or metal-chelating sorbents with bound nickel ions for the isolation of protein complexes with a protein tagged with six histidine residues (6×His).

Similarly to the column chromatography, the optical biosensor methods belong to the group of biochemical approaches. A particularly useful method is the surface plasmon resonance (SPR) effect (Rich & Myszka, 2000); this technique aims at understanding how a protein functions inside a cell by using real time protein dynamics. SPR can detect binding interactions by monitoring the reflection of a

beam of light at the interface between an aqueous solution of potential binding molecules and a biosensor surface carrying an immobilized bait protein. However, optical biosensor methods do not identify the isolated proteins and therefore are combined with mass spectrometric identification (Zhukov, et al., 2004; Grote, et al., 2005).

A frequent problem in cell biology and biochemistry is the identification of a protein or collection of proteins that has been obtained by one of the high-throughput protein separation methods discussed above. Thus, mass spectrometric analysis is a key step in proteomic analysis (Humphery-Smith, et al., 1997). Mass spectrometry is based on the principle that charged particles behave in a very precise dynamics when subjected to electrical and magnetic fields in a vacuum. This allows the separation of charged molecules according to their mass and their charge. Currently, the most commonly used technological platform is the use of the gel-free MudPiT (Multidimensional Protein Identification Technology) involving multidimensional chromatographic separation of proteins and subsequent mass spectrometric analysis. The gel-free MudPIT technology is based on the multidimensional separation of peptides obtained upon the hydrolysis of cell or tissue homogenates (Zgoda, et al., 2009). This approach is widely used due to the simplicity of sample preparation and a high degree of process automation. In some

papers it is referred to as the "shotgun" approach (Gonzalez-Begne, et al., 2009). The eluent from the chromatography stage is directly introduced to the mass spectrometer through electrospray ionization for subsequent mass analysis using MALDI-TOF. In this approach, the proteins that have been isolated previously by using separation techniques are broken into short peptides, ionized with a laser and accelerated in an electric field, and can thus be caught by a detector. MALDI-TOF technique provides a list of masses of the fragmented peptides. Matching this list against a list of pre-calculated peptide masses from an appropriate protein sequence database can help characterizing the isolated protein. Moreover, by employing two mass spectrometers in tandem (MS/MS), it is possible to directly determine the amino acid sequences of the peptides.

All this high-throughput methods have generated a vast amount of interacting data in the last years and there is an interest in combining and comparing these data. However, only a small portion of these protein-protein interactions is detected by more than one method. There are three possible explanations for this: the methods may not have reached saturation, many of the methods may produce a significant fraction of false negatives and false positives, and some methods may have difficulties for certain types of interactions, resulting in complementarities between the methods. Moreover these data are far from being complete in covering protein-

protein interaction networks. For example, previous studies have estimated that 50% of the yeast protein-protein interaction map and only 10% of the human PPI network have been characterized (von Mering, et al., 2002; Hart & Riba-Garcia, 2004). There is a need to develop complementary computational methods in order to bridge the gap.

Computational approaches

Computational approaches for predicting protein-protein interactions can be subdivided into five basic categories: based on genomic information, evolutionary relationships, three dimensional protein structure, protein domains, and primary structure of the proteins (Ivanov, et al., 2010). The different computational approaches are not discussed in this work. Despite the relative success of the computational methods applied to infer protein-protein interactions maps, no approach can accurately predict all protein-protein interactions within an interactome. A number of computational limitations need to be addressed for this to become reality.

Protein-Protein interaction databases

Data management is critical for using high-throughput biological data, including protein-protein interaction data. The massive amount of protein-protein interaction data that have been generated are impossible to handle systematically without a database. To collect, retrieve, and describe protein-protein interactions, several databases have been established and are reviewed in (Chen & Xu, 2003; Tuncbag, et al., 2009). Usually these databases are independent, annotating their own data and developed for specific research interests, using different biological databases as reference. Thus, interaction data is spread across multiple databases and we do not know how much of this information is redundant. The IMEx (The International Molecular Exchange) consortium (Orchard, et al., 2007), which is an international collaboration between different protein-protein interaction databases, has unified efforts, since 2006, to curate protein-protein interaction data according to a standard exchange language (Kerrien, et al., 2007). For this reason Razick, et al. in 2008 have developed iRefIndex, a non-redundant and updated database that provides an index of protein interactions available in 10 primary interaction databases, i.e. BIND (Bader, et al., 2003; Alfarano, et al., 2005), BioGRID (Stark, et al., 2006), CORUM (Ruepp, et al., 2008), DIP (Salwinski, et al., 2004), HPRD (Peri, et al., 2003; Mishra, et al., 2006), IntAct (Hermjakob, et al., 2004; Kerrien, et al., 2007),

MINT (Chatr-aryamontri, et al., 2007), MPact (Güldener, et al., 2006), MPPI (Pagel, et al., 2005) and OPHID (Brown & Jurisica, 2005).

Experimental Structures Provide Protein-Protein Interaction Insights

Proteins interact through their interfaces. There are several fundamental properties that characterize protein interfaces and they can be calculated from the coordinates of the complex. Jones & Thornton (Jones & Thornton, 1996) were the first to characterize the interfaces of four different types of protein-protein complexes: homodimeric proteins, heterodimeric proteins, enzyme-inhibitor complexes and antibody-protein complexes. The complexes were characterized by size and shape, complementarity, residue interface propensities, hydrophobicity, segmentation, secondary structure and conformational changes. As the number of proteins with known structure has grown more groups have addressed the issue of extracting features out of protein complexes.

It has often been assumed that proteins will associate through hydrophobic patches on their surfaces. Again Jones & Thornton found that heterocomplexes are not as hydrophobic as homodimers, since homodimers rarely act independently,

being their interfaces permanently buried. Moreover, transient complexes contain more hydrophilic residues in their interfaces than the permanent complexes. Larsen, et al. in 1998 studied the morphology of protein-protein interactions by analyzing 136 homodimeric complexes. They saw that the interfaces are large, contiguous and form a hydrophobic patch surrounded by a ring of intersubunit polar interactions. The rest of interfaces are characterized by a mixture of small hydrophobic patches, polar interactions and isolated bridging water molecules. Another property that has been studied is the size of the interface. Despite the controversy among studies as for the the size of an interface is concerned, they all agree that transient protein complexes always are smaller than stable ones. In stable complexes, the standard interface size is rather large and confers stability and specificity to the association, on average the size is around 2500 $Å^2$ (Janin, et al., 1988; Janin & Chothia, 1990). In transient complexes, the interface size is smaller compared with that of the stable complex, due to the fact that this interactions are weak and are form or dissociate extremely fast (Lo Conte, et al., 1999; Chakrabarti & Janin, 2002). In another study, Bahadur, et al. in 2003 reported that homodimer interfaces are twice as extensive compared with oligomeric proteins, and are composed by several binding patches.

Proteins interact through their surfaces. Consequently, analyses usually focus on protein surfaces. The determination of which residues and atoms are on the surface

is usually carried out through calculations of the surface area that is accessible to the solvent. Solvent accessibility was firstly introduced by (Lee & Richards, 1971). Usually solvent accessibility it is defined as the resulting surface calculated by rolling a probe sphere center of a given size over the protein's van der Waal surface. Chen & Xu in 2005 and Jones & Thornton in 1997 used the solvent accessibilities to distinguish between interface residues from non-interface ones. They concluded that solvent accessibilities were higher for interacting residues, since non-interface residues tend to reduce their solvent accessibilities by maximizing intra-molecular interactions. Solvent accessibilities can be easily calculated by running NACCESS (Hubbard & Thornton, 1993), an algorithm based on Lee & Richard's idea. Protein surfaces are not flat; rather they are filled with pockets, crevices and indentations (Dundas, et al., 2006). Different approaches have been used to identify cavities and clefts all over the protein surface. For example, Pintar, Carugo, & Pongor have identified protruding and buried residues in proteins developing the CX (Pintar, et al., 2002) and DPX (Pintar, et al., 2003) algorithms. Another powerful tool, SURFNET, generates molecular surfaces and depicts the internal cavities and surface pockets from 3D coordinates of a protein (Laskowski, 1995). These pockets are usually already pre-organized in the unbound state, prior to the protein complexation. Upon forming complexes, protein conformations usually change substantially compared to the unbound protein. At the moment, two main

hypotheses have been formulated in order to explain the bound-unbound transition and have been reviewed by Boehr and Wright in 2008. In the "induced fit" hypothesis (Koshland, 1958) the initial interaction between a protein and its partner produces a change on the conformation. While, in the "conformational selection" hypothesis (Ma, et al., 1999), the unbound protein exists in a number of energetically favorable states. The higher-energy conformation interacts with the binding partner stabilizing the complex.

Additionally, if the molecule interacts with another protein molecule, atoms on the surface of one molecule will interact with atoms on the surface of the partner protein. To understand the nature of the intermolecular interaction several groups have analyzed amino acid propensities in protein-protein interfaces (Jones & Thornton, 1996; Lo Conte, et al., 1999; Glaser, et al., 2001; Zhou & Shan, 2001; Neuvirth, et al., 2004). Jones & Thornton, Lo Conte *et al.* and Neuvirth agree that protein interfaces are characterized by large hydrophobic and uncharged polar residues compared to the rest of the surface when studying heterocomplexes. Specifically, Tyrosine, Methionine, Cysteine and Histidine are the most favored amino acids at the interface. Threonine, Proline, Lysine, Glutamic acid, and Alanine are less commonly found in these regions. It is known that functional residues tend to be highly conserved during evolution. Interface residues are more conserved

compared to non-interface surface residues (Zhou & Shan, 2001; Valdar & Thornton, 2001). There are several computational tools to extract conservation profiles of surface residues; one example is ConSurf (Armon, et al., 2001). Interface residues appear to be less likely to sample alternative side-chain rotamers (Cole & Warwicker, 2002; Liang, et al., 2006), perhaps to minimize entropic cost upon complex formation.

None of the properties described above is by itself sufficient for unambiguous identification of the interface in proteins where the complex with its binding partner is not known. Considerable disagreement exists on which properties are actually useful, and moreover which of them can be combined to increase the power of a prediction method. Many prediction methods have been developed in order to address this problem and have been reviewed by different authors (Zhou & Qin, 2007; de Vries & Bonvin, 2008; Tuncbag, et al., 2009).

Combining Three-Dimensional Protein Structures with Protein Networks

Large-scale high-throughput experimental techniques have produced large amounts of interaction data, characterized by tens of thousands of proteins and potentially hundreds of thousands of relations between them. Thus, abstract representations of the proteome and of the relationships are needed to be able to analyze, manage and interpret such huge collections of data. The proteins can be reduced to a series of nodes that are connected to each other by links, with each link representing a physical interaction between two proteins. The nodes and links together form a network, or, in other words, a protein interaction map.

Many previous studies have explored global aspects of network topology, linking it to protein function, expression dynamics, and other genomic features (de Lichtenberg, et al., 2005; Kelley & Ideker, 2005; Han, et al., 2004; Lee, et al., 2004). In particular, the number of interaction partners or degree is an important factor, and nodes with high degree of interactors, so-called hubs have been found to be essential (Jeong, et al., 2001; Han, et al., 2004). However, most network studies have not considered the structural and chemical aspects of interactions; only a few authors

have proposed to enrich protein networks with structural information of proteins (Aloy & Russell, 2006; Kim, et al., 2006).

We can distinguish between two different kinds of interactions in protein interaction networks (**Figure 1**), mutually exclusive and simultaneous interactions. Simultaneous binding occurs when interactions among the partners can be realized at the same time, using different binding surface regions of the hub and bringing acting and participating proteins together. Examples of simultaneous interactions include large transcription factor complexes, RNA splicing and polyadenilation machinery, protein export and transport complexes.

Conversely, mutually exclusive binding happens when two or more interaction partners can bind to a common or overlapping interface surface region of one protein, being this region only physically available for one partner at any given moment. Mutual exclusive interactions mean that the interaction of one of the partners automatically excludes the occurrence of other interactions. Examples of mutually exclusive interactions include the CDK/cyclin module responsible for cell-cycle progression, the yeast pheromone response pathway, MAP signaling cascades, etc.

By relating known three-dimensional structures to protein-protein interaction networks it is possible to determine which of the multiple interactions or connections that are made to a common binding partner can occur simultaneously, and which are mutually exclusive due to overlapping binding surfaces (**Figure 1**) and, consequently, to make inferences about which parts of the protein surfaces are involved in the various interactions.

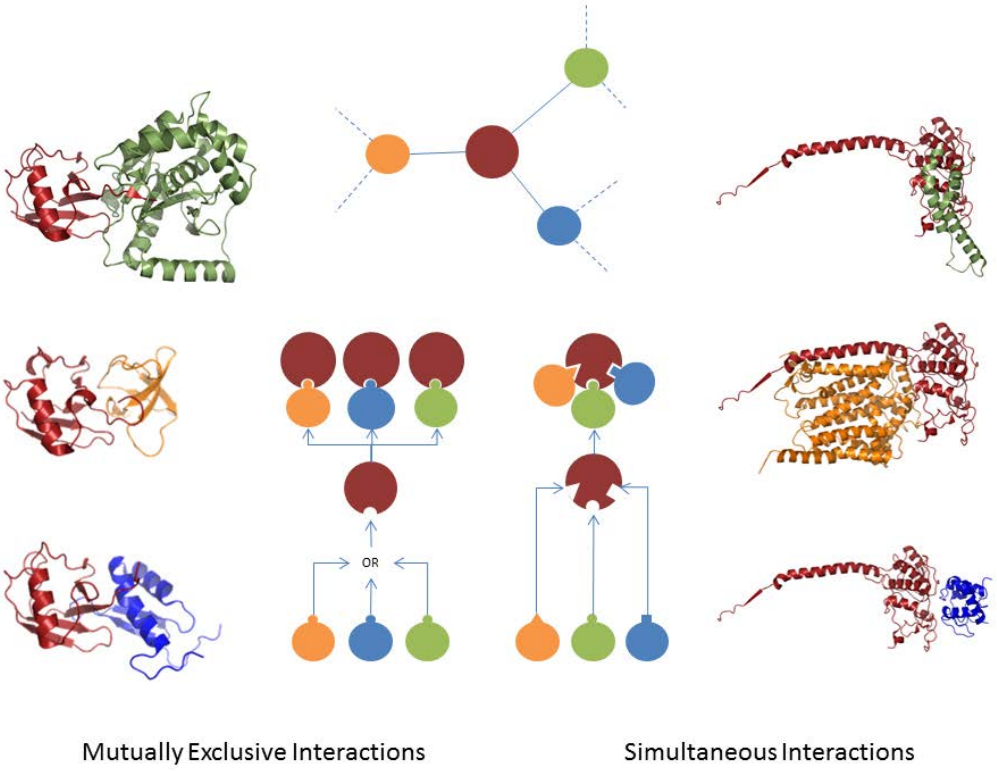


Figure 1: Simultaneous versus mutually exclusive interactions. On top a simplified network diagram is represented, a hub or central protein (red) connects three other proteins (green, yellow and blue nodes). Connections with the hub can occur simultaneously (right) or in a mutually exclusive fashion (left). On the left, simultaneous interaction examples, the cytochrome b-c1 complex: the hub, in red, is the CY1_YEAST (PDB: 3CX5, chain: O), which interacts with QCR6_YEAST (PDB: 3CX5, chain: Q) in green, CYB_YEAST (PDB: 3CX5 chain: N) in yellow and CYC1_YEAST (PDB: 3CX5

chain: W) in blue. On the right, mutually exclusive examples of ubiquitin partners: the hub, in red, is the UBIQ_HUMAN), which interacts with UCHL3_HUMAN (PDB: 1XD3, chain: B) in green, SH3K1_HUMAN (PDB:2K6D chain: N) in yellow and PLAP_HUMAN (PDB: 2K8B chain: W) in blue.

Aim and Contributions of the Study

Proteins are responsible for an impressive large variety of functions. To properly understand the significance of protein-protein interactions in the cell it is important to address two problems: first, is identification of the different interactions that are involved in each biological function, and, second, is determining how proteins interact and the consequences of the interaction.

The identification of protein interactions by high-throughput experiments has led to the development of a number of methods for their analysis, producing, in the last years, a vast amount of interacting data. However, there are at least two issues that arise from the analysis of such experimental maps, these are, on one side, the significant number of false positives they contain and, on the other, the difficulty in distinguishing whether, when more than one protein interact with the same partner, they can do so simultaneously, i.e. whether their interaction is mutually exclusive.

The aim of the present study is to combine known three-dimensional structures with protein-protein interaction networks to determine which of the multiple interactions or connections that are made by a hub can occur in mutually exclusive fashion, and, in such cases, identify, whenever is possible, the shared similarities in their binding regions, concluding that their interaction has to be mutually exclusive (i.e. not simultaneous) and that the region identified by similarity is indeed the interaction site.

Materials and Methods

Protein-Protein Interaction Data

Protein-protein interaction data are obtained from iRefIndex (Razick, et al., 2008), release 7.0 (May 18th 2010), a non-redundant and updated database, that provides an index of protein interactions available in 10 primary interaction databases, i.e. BIND (Bader, et al., 2003; Alfarano, et al., 2005), BioGRID (Stark, et al., 2006), CORUM (Ruepp, et al., 2008), DIP (Salwinski, et al., 2004), HPRD (Peri, et al., 2003; Mishra, et al., 2006), IntAct (Hermjakob, et al., 2004; Kerrien, et al., 2007), MINT (Chatr-aryamontri, et al., 2007), MPact (Güldener, et al., 2006), MPPI (Pagel, et al., 2005) and OPHID (Brown & Jurisica, 2005). Among the available interactomes, I have selected the ones of the species *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorabditis elegans*, *Escherichia coli* and *Saccharomyces cerevisiae*.

I filtered interactions in order to collect only intraspecific binary interactions (i.e. where both partners belong to the same species), where both contributors are annotated in UniProt database (Jain, et al., 2009; The UniProt Consortium, 2011).

Three dimensional data: The Protein Data Bank

The PDB archive is a repository of atomic coordinates of experimentally determined structures. By the end of the 1980's the number of structures started to increase dramatically (**Figure 2**) and that growth continues to date (Bernstein, et al., 1977), as well as the complexity of the structures that has been deposited yearly the complexity of the structures that could be determined grew, making possible to solve virus structures (Arnold & Rossmann, 1988) or even larger structures, including molecular machines such as the ribosome (Moore, 2001).

The primary information stored in the PDB consists of coordinate files for biological molecules. These files are list of atoms for each protein and their 3D localization (X, Y and Z coordinates) in space. Half of these files are protein complexes obtained with X-ray crystallography or Nuclear Magnetic Resonance

(NMR) and represent an important source of information at a high level of detail to study the molecular bases of protein-protein interactions, and more generally of protein complex formation (Levy, et al., 2006). Unfortunately, the complexes in the PDB are also highly redundant. The database has an inherent bias towards certain complexes such as antibody-antigen or enzyme-inhibitor complexes while others, such as membrane complexes, are underrepresented (Ezkurdia, et al., 2009).

The protein-protein interaction data was enriched with structures obtained by X-ray crystallography or NMR techniques found in the PDB (Bernstein, et al., 1977), in order to obtain atom spatial coordinates to which could be applied this analysis. Whenever more than one PDB accession code is associated to a protein, I select the one obtained by X-ray crystallography, covering as much as possible the chain sequence and with the highest resolution.

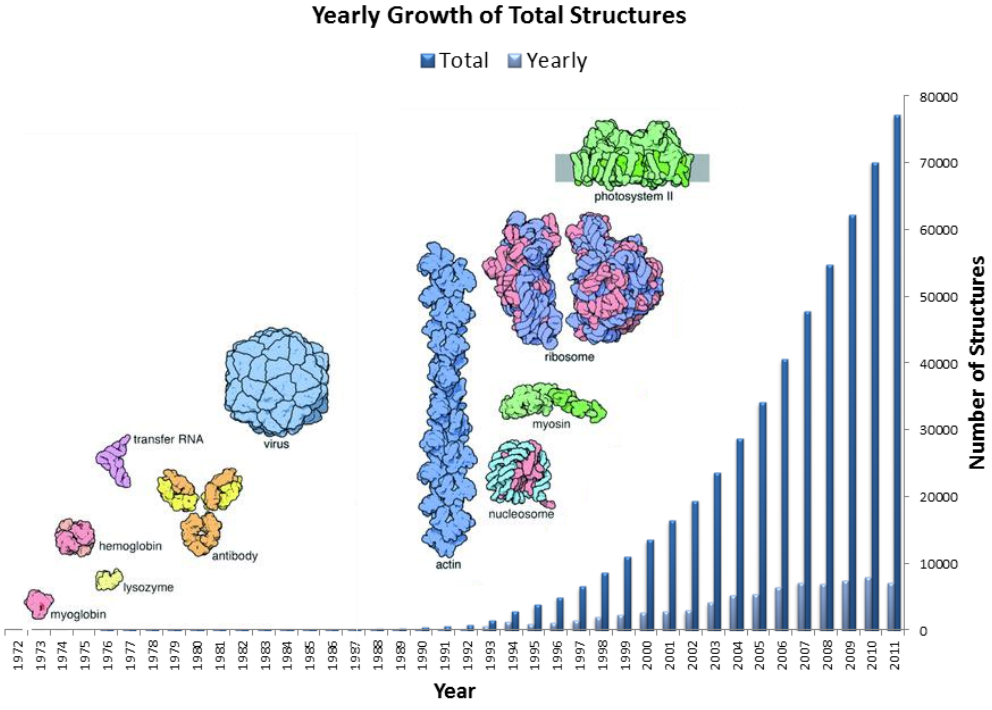


Figure 2: The growth of the number of structures in the PDB archive 1972-2006 (Modified from Berman, 2008).

Sub-network Definition and Analysis

Protein-protein interactions are usually represented by networks, in which nodes represent proteins and edges represent experimentally observed interactions. An interactome-representing network can be divided in a set of sub-networks where a central node is surrounded by adjacent nodes directly connected to it by an edge. The central node is a specific protein and the adjacent nodes are its partners of interaction. Since the purpose is to identify which interactors are mutually exclusive, and in such cases determine the conserved structural exposed regions among binders of the same protein, I have considered only those sub-networks with at least 3 partners with known structure. The sub-networks so defined represent the principal element of the analysis.

A redundancy filter was applied to all interactors within a particular sub-network using PISCES (Wang & Dunbrack, 2003). PISCES is a standalone package for culling sets of protein sequences from the Protein Data Bank (PDB) by sequence identity and structural quality criteria, using the method of Hobohm and Sander (Hobohm, et al., 1992; Hobohm & Sander, 1994). I have defined redundant pairs of protein structures as those pairs that share more than 30% sequence identity. Among

redundant pairs I selected those with best structure according to the structural quality criteria, i.e. experiment type (X-ray or NMR), resolution, and R-value.

Identification of Local Structural Similarities

In order to fin solvent exposed structure similarities among interactors within a sub-network, I evaluated the solvent accessibilities with NACCESS program (Hubbard & Thornton, 1993). NACCESS is the implementation of Lee and Richards's method (Lee & Richards, 1971). It calculates the atomic surface of a protein chain defined by rolling a probe of a given size around the van der Waals surface. I have run the program using default parameters: a probe size of 1.40 Angstroms and the van der Waal radii from (Chothia, 1976). I defined as solvent accessible residues those residues that are 50% accessible compared to the accessibility of that residue type in an extended ALA-x-ALA tripeptide, discarding those residues under the threshold and modifying the PDB files to only contain solvent accessible residues. The use of surfaces highly reduces the number of structural comparisons, speeding up the identification of local similarities. I then submitted the surface regions of the sub-network interactors to FunClust (Ausiello, et al., 2008) a publicly available tool consisting of a two-step procedure.

In the first step, the Query3D algorithm (Ausiello, et al., 2005) identifies all the pairwise similarities among the chains within the sub-network. Query3D is an unsupervised structure comparison method that searches for the largest subset of matching amino acids between two protein chains. Matching amino acids requires three criteria to be fulfilled: the residues must be neighbors in the space, and they must share a structural and biochemical similarity. Two sets of amino acids are considered structurally similar when they locally superpose within an RMSD threshold that was set to 2.1 Å. The RMSD score is calculated using two points per amino acid: one is the C-alpha atom and the other is the geometric average of the side-chain atom coordinates. The algorithm uses the DayHoff's substitution matrix as default measure to evaluate the biochemical similarity. The matrices proposed in Dayhoff, et al. in 1978 are based on the concept of PAM (point accepted mutation) and are called PAM matrices. An accepted point mutation in a protein is a replacement of one amino acid by another every 100 amino acids. Query3D match two amino acids if their similarity according to the matrix PAM250 is between 0.3 and 1.2.

In the second step a clustering algorithm represents all the pairwise similarities as nodes of a graph, connecting them when the corresponding chains also share a group similarity, therefore identifying clusters of chains with a local

structural similarity as connected paths in the graph. The clusters are sorted by an approximate significance score, called FunClust score, calculated by multiplying the number of residues in the group similarity by the number of chains belonging to the cluster (Ausiello, et al., 2008).

Superposition and Conservation Scores

I further scored the FunClust results using the LGA package (Zemla, 2003). The scores used to measure superposition between similar residues among different proteins within the same sub-network, described in more detail below, were: the overall root mean square deviation (RMSD) of all corresponding C-alpha atoms, the global distance test (GDT-TS), and LGA scores, that not only calculates a 'best' superposition between two proteins, but also identifies the regions of local similarity between compared structures. Besides, as a result of LGA processing, I obtain the rotated coordinates for all structures when compared with one structure taken as reference. Next I calculate the average of the scores mentioned before for each cluster, as well as the average conservation score.

RMSD (Root Mean Square Deviation) is a measure that calculates the structural divergence between two protein structures and is defined as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(x_i - x_i^{'}\right)^2 + \left(y_i - y_i^{'}\right)^2 + \left(z_i - z_i^{'}\right)^2\right]}$$

being (x_i, y_i, z_i) and (x_i', y_i', z_i') the coordinates of the corresponding atoms that I want to superimpose to each other and N the number of atom pairs which are compared. In this case the correspondence between the pairs of atoms I want to superimpose is known. As a result of LGA processing I obtained the rotated coordinates for all structures when compared with another structure taken as reference. The rotated coordinates were obtained by applying the rigid-body translation T = (T_x, T_y, T_z) and rotation R = (R_x, R_y, R_z) to one of the proteins that minimizes the RMSD between the given set of atom pairs:

$$RMSD(T,R) = min_{T,R}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(x_i - R_x x_i^{'} + T_x\right)^2 + \left(y_i - R_y y_i^{'} + T_y\right)^2 + \left(z_i - R_z z_i^{'} + T_z\right)^2\right]}$$

GDT-TS (Global Distance Test - Total Score) is a measure that performs sequence-independent superposition of two given structures and calculates the number of structurally equivalent pairs of C-alpha atoms that are within a specified distance. The GDT-TS score is calculated as follows:

$$GDT - TS = 100 * \frac{\sum_{d_i} \frac{GDT_{d_i}}{NT}}{4} \qquad d_i \in \{1.0, 2.0, 4.0, 8.0\}$$

and is the average of four scores obtained with four different distances, d = 1.0, 2.0, 4.0 and 8.0 A, divided by the number of residues of the target (NT). The GDT-TS is one of the standard measures used in CASP, Critical Assessment of Techniques for Protein Structure Prediction (Zemla, et al., 1999).

The LGA_S (Zemla, 2003) can be defined as a combination of GDT and RMSD values and can be used to evaluate the level of structural similarity of selected regions. The LGA_S is a two component scoring function: LCS (Longest Contiguous Segments) and GDT (Global Distance Test), described previously, defined by the following formula:

$$LGA\_S = w * S(GDT) + (1 - w) * S(LCS)$$

where w is a parameter w ($0.0 \leq w \leq 1.0$) representing a weighting factor, and S(F) is a function that is defined as follows:

$$foreach\ vi(v1, v2, \dots, vk)\left\{Y = \frac{(k - i + 1)}{k}; X = X + Y * F_{vi};\right\}$$

$$S(F) = \frac{X}{\left((1+k)*\frac{k}{2}\right)};$$

The LGA_S score is calculated with reference to the number of residues in the target protein. It ranks from 0 to 100, where lower values indicate less similar regions.

The LGA_Q (Zemla, 2003) which means LGA quality score is calculated with use of the formula:

$$LGA\_Q = \frac{0.1 * N}{0.1 + RMSD}$$

FunClust Score: The score is given by the number of residues in common between all the matches multiplied by the number of matches belonging to the cluster (see Identification of Local Structural Similarities).

Conservation score: when I compare two amino acids they can be either identical or different, in other words, can have more or less similar chemical properties. There is a need to quantify these similarities or differences; this is to estimate the probability that one amino acid is replaced by another during evolution. These values are empirically derived and reported in tables called similarity or substitution matrices. In these matrices, each row and each column corresponds to 1 of the 20 amino acids, and each cell contains a measure of the probability that the amino acids in the column and in the row can replace each other during evolution. The BLOSUM, blocks substitution matrix, matrices are derived by use of local alignments of well conserved regions in homologous proteins (Henikoff & Henikoff, 1992). I have used BLOSUM-30 matrix, derived from alignments sharing more than 30% identity with any other sequence in the alignment.

Identification of Overlapping Interfaces

Statistical methods

To estimate how well Estrella procedure performs I calculated, for each one of the clusters for which I know the answer (see Results), the True Positives, True Negatives, False Positives and False Negatives as follows:

- The True Positives (TP) are the mutually exclusive interactors that Estrella identifies correctly;
- The True Negatives (TN) are all the interactors that are not mutually exclusive and Estrella identifies correctly;
- The False Positives (FP) are all the interactors that Estrella identifies as mutually exclusive, while they are not;
- The False Negatives (FN) are all the interactors that Estrella does not correctly identify and they actually are mutually exclusive.

The diagnostic power of a method can be expressed in terms of Sensitivity and Specificity, or Positive and Negative Predictive Values:

Sensitivity (Se): measures the proportions of positives, mutually exclusive interactors, that are correctly identifies as such. In other words, Sensitivity, estimates the ability of a method to find the positive cases.

$$Sensitivity(Se) = \frac{TP}{TP + FN} \times 100$$

Specificity (Sp): measures the proportions of negatives, not mutually exclusive interactors, which are correctly identify. This is the ability of a method to detect negative cases.

$$Specificity(Sp) = \frac{TN}{TN + FP} \times 100$$

Positive Predicted Value (PPV) or precision rate is the proportion of mutually exclusive interactors that are correctly detected. It reflects the probability of a method to correctly identify True Positives.

$$Positive\ Predictive\ Value\ (PPV) = \frac{TP}{TP + FP} \times 100$$

Negative Predictive Value (NPV) is the proportion of non-mutually exclusive interactors that are correctly detected. It reflects the probability of a method to correctly identify True Negatives.

$$Negative\ Predictive\ Value\ (NPV) = \frac{TN}{TN + FN} \times 100$$

The Accuracy combines the previously mentioned concepts.

$$Accuracy\ (Ac) = \frac{TP + TN}{TP + FP + FN + TN} \times 100$$

Results

The Estrella Project

The Estrella project is designed to characterize mutually exclusive interactors in terms of local protein surface similarities among a non-redundant set of proteins sharing a common interacting partner. I applied it to seven different organisms (*H.sapiens*, *R.norvegicus*, *M.musculus*, *D.melanogaster*, *C.elegans*, *S.cerevisiae*, *E.coli*), making all the pre-computed data available via a web server. The following steps represent a summary of the Estrella pipeline (see and Materials and Methods):

- Collection of protein-protein interactions belonging to a given species where both partners are annotated in UniProt (**Table 1**). Construction of protein-protein interaction maps;
- Incorporation of protein structures to the protein-protein interaction map, when available;

- Extraction of sub-networks (for the sub-network definition, see Material and Methods);
- Assessment of non-redundant protein structures for each sub-network;
- Calculation of solvent accessible surfaces for proteins of known 3D structure;
- Comparison of all against all solvent accessible surfaces to retrieve sets of at least three exposed residues that are structurally similar.
- Evaluation of the structurally similar sets in terms of superposition quality indexes and conservation score;
- Storage of interaction and result data in the Estrella database.

Interactome Analysis

The identification of mutually exclusive interactions is possible solely when experiments allow the determination of the binding site of the common partner o with two interactors. If the binding site is unique or at least partially overlapping, then the two interactors can be considered as mutually exclusive. Such type of information is not always available and a reliable validation is possible only when

experimentally determined structures of the complexes between the central protein or hub and more than one of its interactors are known. The coverage of such knowledge on protein-protein interaction data is rather scarce, therefore it is not a simple task to estimate the number of mutually exclusive interactions in a cell.

Applying the Estrella procedure to the seven interactomes I retrieved a total amount of 283227 intraspecific pairwise interactions annotated by UniProt, involving a total number of 37406 unique proteins. Out of these proteins only 12.36% have a 3D structure available in the PDB (**Table 1**). Among the species analyzed, this percentage varies substantially, ranging from 1.54% for fruit fly to 36.55% for *E.coli*. The PDB structure coverage for human proteins involved in analyzed interaction is around 20%.

Out of the complete collection of sub-networks, i.e. sets of more than three proteins interacting with the same hub, 8817 contain at least three non-redundant proteins of known structure and could therefore be analyzed with the procedure described above. In 7310 cases, I could identify the presence of structurally similar regions in proteins interacting with the same hub, which are candidates for being mutually exclusive interactors (**Table 2**).

Species	Number of pairwise interactions	Number of proteins involved in the interaction	Number of involved proteins of known structure
H.sapiens	72398	12294	2451(19.94%)
S.cerevisiae	157257	6023	649(10.78%)
D.melanogaster	36629	9570	147(1.54%)
M.musculus	3904	3052	340(11.14%)
C.elegans	10382	4934	838(16.84%)
R.norvegicus	1047	838	46(5.49%)
E.coli	1610	695	254(36.55%)
Total	283227	37406	4725(12.36%)

Table 1: Datasets used for the analysis.

Species	Sub-networks		Presence of structurally common regions	
	all	nr	all	nr
H.sapiens	5176	4598	3983	3721
S.cerevisiae	3971	3796	3446	3362
D.melanogaster	156	137	45	39
M.musculus	65	46	26	16
C.elegans	13	12	9	9
R.norvegicus	78	63	27	23
E.coli	171	165	142	140
Total	9630	8817	7678	7310

Table 2: Sub-networks extracted, sets of more than three proteins interacting with the same hub and sub-networks where I could identify the presence of structurally similar regions, which are candidates for being mutually exclusive interactors. nr: non-redundant.

Identifying Mutual Exclusive Interactions

To estimate the accuracy of the method, I extracted from the dataset all cases (152 sub-networks) where an experimentally determined structure of the complexes between the central protein and more than one of the *bona fide* mutually exclusive interactors are known. In such cases, I identified contact residues in the interface of the complex by running the Atom Nucleus Distance under PIA, a subprogram contained in PSAIA software (Mihel, et al., 2008). If the binders interact with at least 3 common residues of the central protein, I considered them as mutually exclusive interactors (128 sub-networks). I submitted the examples to Estrella following the same procedure described before in Material and Methods.

Let us assume that there is a sub-network where a central protein interacts with M + N proteins where M are experimentally known to interact with the same region of the central protein and N are not and that, for the same sub-network, Estrella produces a cluster of m proteins predicted to establish mutually exclusive interactions. As defined in material and methods, the True Positives (TP) are M ∩ m; the False Positives are m-M; the True Negatives are N ∩ n and the False

Negatives are n-N. In other words, for each cluster, I count how many times I detect the correct mutually exclusive interactions (True Positives), how many times I include in the set of mutually exclusive proteins some that are not (False Positives), how many times I miss a mutually exclusive interaction (False Negatives) and, finally how many times I correctly predict that a protein of the sub-network does not bind to the same surface of the hub as the others in the sub-network.

Figure 3 schematizes the definition of these parameters in more complex cases. In the left upper part of the figure I show the experimentally known situation where A1, A2 and A3 interact with the same region of the hub, the interaction of B1, B2 and B3 with the hub is also mutually exclusive, although they bind to a region different from that of the As. C1 binds to a region different from both the A and B binding sites. The example represents a possible set of sub-networks predicted as mutually exclusive by Estrella and the corresponding values for FP, TP, TN, FN, specificity (Sp) and sensitivity (Sn). The overall values for the specificity and sensitivity are computed as the average of the values for each identified cluster.

As it can be appreciated from **Table 3**, the method has an accuracy above 77%, with a higher specificity (85%) than sensitivity (about 70%). It should also be

mentioned (**Table 4**) that rarely I fail to identify more than one partner (less than 0.1% of the cases), while more often the prediction includes one protein that in reality does not establish a mutually exclusive interaction.

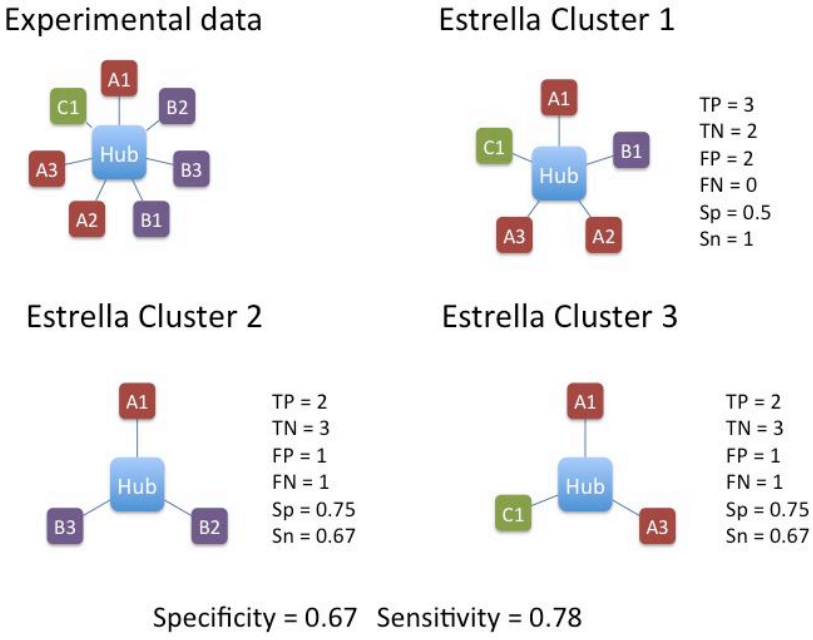


Figure 3: Exemplification of the way I compute the statistical parameters.

This is, in my opinion, relevant, since it can direct the design of a limited number of experiments to validate the hypothesis. The sensitivity increases when

only the first ranking cluster is considered at the expense of a 20% decrease in specificity. The overall accuracy is very similar in the two cases.

		Mutually Exclusive Interactors		
		Positive	Negative	
Estrella	Positive	**TP** 4428 (260)	**FP** 878 (95)	**Positive Predictive Value** 83.45% (82%)
	Negative	**FN** 1898 (36)	**TN** 5162 (57)	**Negative Predictive Value** 73.12% (72%)
		Sensitivity 70.00% (88%)	**Specificity** 85.46% (63%)	**Accuracy** 77% (79%)

Table 3: Statistical parameters for the Estrella method applied to the sub-networks where the experimental structures of complexes between the hub protein and at least two partners are available. Data are computed as the average of all clusters for each sub-network and only considering the first ranking clusters (between parentheses).

Clusters	%
With more than one missing partner	8.72
With one missing partner	40.4
Perfectly defined	50.5
With one extra partner	0.23
With more than one extra partner	0.06

Table 4: Results of the Estrella procedure applied to sub-networks for which the experimental structure of the complexes is known. Data are shown for all clusters.

The identification of the common substructures often provides a correct prediction of the node binding sites as well. As shown in **Table 5**, I correctly identify 26% of the residues that are indeed buried in the complex interface on average. The figure rises to 31% if only the first ranking cluster is considered. Furthermore, I am able to correctly predict at least one interface residue in 63% of the cases (75% for the first ranking clusters) (**Table 4**). This is relevant since it might help reducing the search space in docking algorithms.

	All clusters	First ranking cluster
Number of correctly predicted common interfaces complexes	1739	89
Total number of residues at the interface	34306	976
Number of correctly identified interface residues	9192	300
Number of common interfaces where at least one interface residue is correctly identified	1101	67

Table 5: Number of correctly identified interface residues in the correctly identified complexes

The Estrella Database and Web Interface

Estrella Database Design and Construction

Based on the relational model for database management (Codd, 1998), the Estrella database consists of 14 tables: ACCESSION_NUMBER, CLUSTER, GENE_NAME, LOCI, GO, GO_TO_PROT, INTERACTOME, LGA,

ORGANISM, PROTEIN, REFSEQ, SIMILARITIES, STRUCTURE, SUBNETWORK (**Figure 4**).

The ORGANISM table contains all the interactomes used in the analysis (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorabditis elegans*, *Escherichia coli* and *Saccharomyces cerevisiae*) with each acronym name and taxonomic identifier, as well as the total number of binary interactions, binary interactions where both participants are annotated in UniProt database (Jain, et al., 2009; The UniProt Consortium, 2011), proteins and known structures per organism.

For each organism, I collected all binary interactions in the table INTERACTOME grouped by sub-networks (named as the central protein or hub) and annotating whether an interactor is part of the "all interactors" or "non-redundant interactors" network (see Material and Methods, The Estrella Project section).

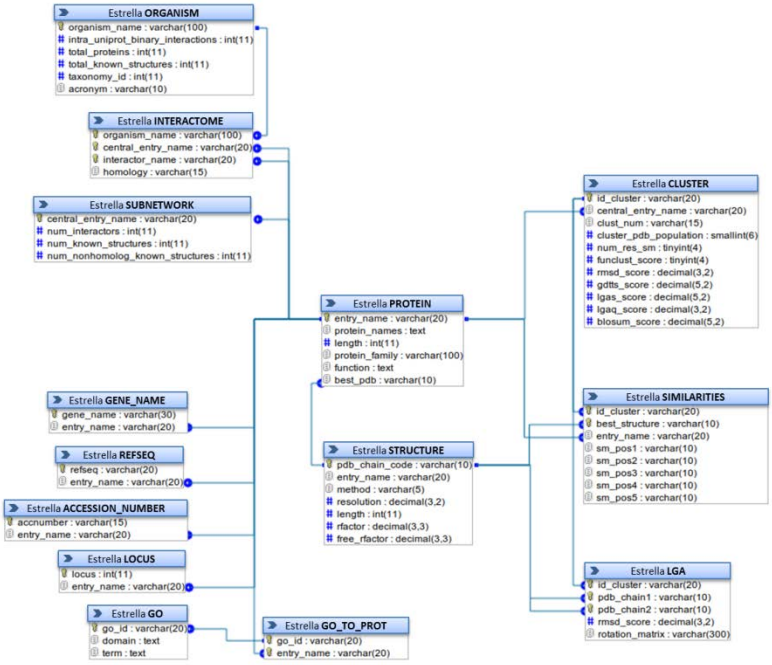


Figure 4: Visual overview of the Estrella database and the relations (connectors) between its tables.

All sub-network features are summarized in the SUBNETWORK table, so that each sub-network has the total number of interactors, associated known structures and non-redundant known structures. Clusters of proteins shearing the same local structural similarities within the same sub-network are contained in the table CLUSTER. For each cluster an index is generated (**Figure 5**), taking into

account the name of the sub-networks' central protein, the cluster ranking provided by FunClust and the approach, all or non-redundant. For each cluster features such as number of proteins, similar residues within the cluster and superposition measures (FunClust score, RMSD, GDT-TS, LGA s, LGA q and conservation score) are given.

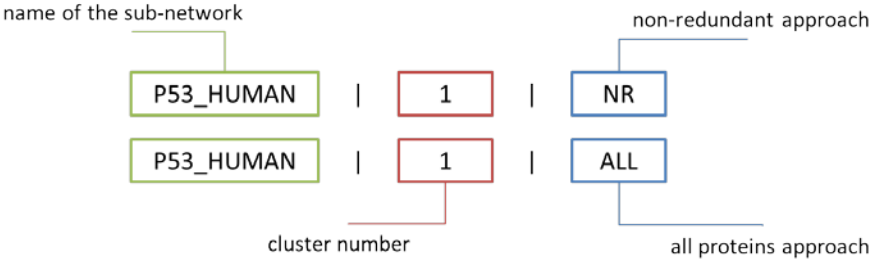


Figure 5: Generation and assignment of indexes in the Estrella Database. On the top two examples that explain how the indexes are generated for each one of the clusters stored in the database are shown. Each index is composed by three fields separated by linkers"|": name of the the sub-network, cluster number and approach.

The SIMILARITIES table contains the UniProt entry name, the selected PDB chain and the identified similar residues ordered by their position in every interactor within each cluster.

Rotation matrices and translation vectors needed to superimpose structures given the local structure similarities are specified in the table LGA for all against all interactors within each cluster and for all clusters.

The PROTEIN table contains protein's length, family, function and selected PDB chain, when known. PDB chain features (method, resolution, length, rfactor and free rfactor) are stored in the table STRUCTURES. Gene Ontologies terms and domains are stored in GO table and are linked to the PROTEIN table through the GO_TO_PROT table.

The ids for each protein in the database, RefSeq number/s, protein name/s, locus/loci gene name/s or UniProt accession number/s are associated to the PROTEIN table by the UniProt entry name and are located respectively in REFSEQ, ENTRY_NAME, LOCI, GENE_NAME, AND ACCESSION_NUMBER tables.

Estrella Web Interface

The Estrella database and web server follow a solution stack of open source software whose components are: Linux (operating system), Apache HTTP Server, MySQL (database software), and PHP programing languages. The Web interface was developed in the HTML language with embedded scripts in PHP, Perl and

JavaScript and Cascading Style Sheets (CSS), easy to interpret and navigate. It is hosted on an Apache web server at (http://bl210.caspur.it/ESTRELLA/home.php).

Search Features: As mentioned above, the results obtained for the seven analysed interactomes are stored in the Estrella database. The database can be searched both with a "Protein of Interest" or an "Organism". Searches by "Protein of Interest" may be based on a wide range of supported identifiers, including gene name, UniProt entry name, RefSeq number and UniProt identity code. All sub-networks including the protein will be listed, as well as the sub-network where the protein of interest is the central protein of the sub-network. Search by "Organism" will provide the full list of sub-networks for each of the seven organisms collected in the database, sorted by descending number of known structures.

By selecting one of the listed proteins, either coming from the organism or protein searches, the user navigates to a sub-network information page (**Figure 6**). On the top of it, a dropdown box named "General Information for Protein" (**Figure 6c**) contains the central protein features and crosslinks to other databases: CATH (Orengo, et al., 1997), PDB (Bernstein, et al., 1977), UniProt (Jain, et al., 2009; The UniProt Consortium, 2011), iRefIndex (Razick, et al., 2008), SCOP (Murzin, et al.,

1995), NCBI (http://www.ncbi.nlm.nih.gov), and Gene Ontologies (Ashburner, et al., 2000). Two buttons in the Results box (**Figure 6d**) link to the page result, depending on whether the user is interested in the local structural similarities found in the whole set of binders ("All Interactors" button) or in the subset of non-redundant binders ("Non-Redundant Interactors" button). The Interactor Information box shows an interactive sub-network scheme (**Figure 6e**) via a Cobweb applet (von Eichborn, et al., 2011), where nodes and edges are differentially colored on the basis of the 3D structure knowledge or of the redundancy between chains. Moreover, nodes having a known structure are represented with a picture of their selected PDB chain (see Materials and Methods), when it was available in the PDBsum database (Laskowski, et al., 1997) . A table below (**Figure 6f**) lists all the binder nodes with their UniProt name, protein name, family, selected PDB chain and redundancy if the structure is known.

While the user is surfing Estrella in Search mode, a left panel appears, the "History Index" (**Figure 6b**), registering all the steps, allowing backward and forward navigation through page linking buttons. If a sub-network is selected and structural similarities are found, a button will appear offering the possibility to download all the data.

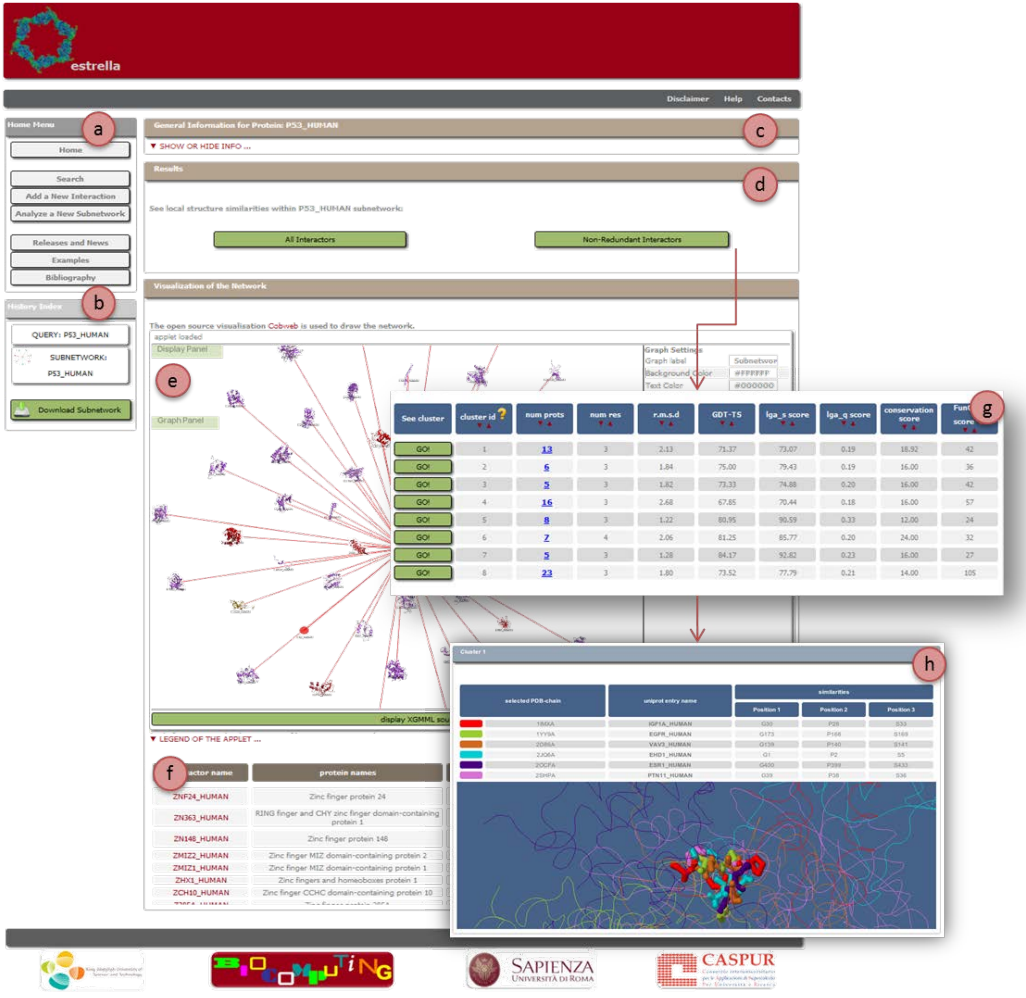


Figure 6: Snapshot of the Estrella Web Server showing the sub-network and the result page.

Extra Functionalities and Applications**:** The Estrella web server gives two extra possibilities: adding a new interaction to an existing sub-network, thus evaluating how the known local similarities (if existing) change by repeating the analysis on the modified sub-network, and submitting a completely new interaction sub-network in order to identify the local structure similarities among the binders that it contains. The usage of these applications, which do not imply the update of the database, is obtained via the corresponding buttons "Adding a new interaction" and "Analyzing a new sub-network" in the left command panel of the server (**Figure 6a**).

The need of Adding a New Interaction is evident whenever it is possible to enrich a sub-network that is already present in the database with new data provided by the user. The input form requires two interacting proteins of known 3D structure in the PDB, with at least one of them already present in the Estrella database (which identifies the sub-network that will be modified). The information already contained about the submitted proteins are quickly visualized, allowing the user to examine the known PDB structures in the database and eventually to upload a user-selected PDB file not shown in the page. The final submission starts the algorithm that identifies the surface similarities on the modified sub-network. A status-log will appear informing the user on the running progress. If both proteins are present in Estrella,

the analysis will present results for the corresponding sub-networks, provided that both contain a sufficient number of PDB structures. As when browsing the Estrella database, the user can access the results either for the whole sub-network or for the non-redundant binders belonging to it.

Estrella can be used for real time analysis of a new interaction sub-network. The user can input a set of PDB files, the identifier of each binder chain and the accessibility threshold to evaluate the corresponding surface residues. The routine automatically checks the correctness of the input file size and format. It must be noticed that in this application the algorithm will analyze the whole set of chains to identify the surface similarities, therefore the user should be aware of possible homology relationship among the submitted protein chains.

Result pages: In both cases, Search or Add features, the result page are always structured in the same fashion. The result page contains a sortable table with all clusters found by the Estrella procedure (**Figure 6g**), allowing the user to select the cluster of interest ranked by FunClust score, number of similar residues, number of proteins, average RMSD, GDT-TS, LGA or conservation scores. All proteins grouped in each cluster will appear, in a popup window showing an interactive

representation of the proteins and residues involved in such cluster via the Cobweb applet (von Eichborn, et al., 2011) by clicking in the cell corresponding to the number of proteins. After the cluster selection (clicking in the Go! button), a popup window will appear showing a table with the proteins and the identified similar residues, organized by position (**Figure 6h**). A Jmol applet (http://jmol.sourceforge.net) will then show the best local structure superposition for the selected cluster.

Examples

Case study Ubiquitin

Although ubiquitination is often a signal for degradation by the proteasome (Hershko & Ciechanover, 1982), it has become clear that this modification in fact leads to a variety of responses like DNA repair, cellular trafficking, immune responses, and chromatin remodeling (Petroski, 2008).

Ubiquitin is a small protein of 76 amino acids with a small surface area. Over twenty distinct ubiquitin-binding domain (UBD) families and more than 200 interactors have been identified (Dikic, et al., 2009). In the Estrella database, 139

known structure interactors are stored, 108 of which are non-redundant. In PDB I found 15 experimentally determined structures containing ubiquitin and one of its interactors, (**Table 6**) which allows a precise identification of the binding interface.

PDB Code	Ubiquitin Chain/s	Interactor Chain/s	UniProt Entry Name/s	Interactor Name
1NBF	C|D	A|B|E	UBP7_HUMAN	Ubiquitin carboxyl-terminal hydrolase 7
1S1Q	B|D	A|C	TS101_HUMAN	Tumor susceptibility gene 101 protein
1XD3	B|D	A|C	UCHL3_HUMAN	Ubiquitin carboxyl-terminal hydrolase isozyme L3
2DEN	B	A	SB132_HUMAN	Ubiquitin-like protein 7
2FUH	A	B	UB2D3_HUMAN	Ubiquitin-conjugating enzyme E2 D3
2HTH	B	A	VPS36_HUMAN	Vacuolar protein-sorting-associated protein 36
2IBI	B	A	UBP2_HUMAN	Ubiquitin carboxyl-terminal hydrolase 2
2K6D	B	A	SH3K1_HUMAN	SH3 domain-containing kinase-binding protein 1
2K8B	A	B	PLAP_HUMAN	Phospholipase A-2-activating protein
2KDF	B|C	A	PSMD4_HUMAN	26S proteasome non-ATPase regulatory subunit 4
2KHW	B	A	POLI_HUMAN	DNA polymerase iota
3IFW	B	A	UCHL1_HUMAN	Ubiquitin carboxyl-terminal

				hydrolase isozyme L1
3IHP	C|D	A|B	UBP5_HUMAN	Ubiquitin carboxyl-terminal hydrolase 5
3JW0	X|Y	C|D	NED4L_HUMAN	E3 ubiquitin-protein ligase NEDD4-like
3LDZ	E|F|G	A|B|C|D	STAM1_HUMAN	Signal transducing adapter molecule 1

Table 6: Experimentally determined commplexes containing ubiquitin. By column: PDB entry, ubiquitin chain/s, interactors chain/s, UniProt accession number, interactors UniProt entry name and protein name.

When superposing all the complexes using as reference the ubiquitin, all proteins show overlapping interfaces except for 1S1Q_A and 3JW0_A that I excluded from this analysis. Moreover, chain A of the complex 3IHP represents an ubiquitin interactor (the Ubiquitin carboxyl-terminal hydrolase 5), 854 residues long, that covers ubiquitin surface almost entirely and involving several contact residues and I excluded it from this study. The remaining 12 interacting chains were considered for the analysis.

I obtained a clear example of locally similar surface regions in cluster 1, according to FunClust score. Cluster 1 (**Figure 7)** consists of two Serines (S) in position 1 and 2, and a hydrophobic residue in position 3: Methionine (M), Leucine

(L) or Phenylalanine (F). It groups together eight different proteins (1XD3C, 2FUHA, 2HTHB, 2IBIA, 2KDFA, 2K8BB, 3IFWA and 3LDZD) six of which show the similarity region in close proximity of the binding interface with the central protein.

Case study SUMO1

The SUMO1 family members are expressed throughout the eukaryotic kingdom. Despite sharing only 18% sequence identity with ubiquitin, human SUMO1 (small ubiquitin-related modifier) possesses the characteristic ubiquitin-fold common to ubiquitin-like proteins (Mayer, et al., 1998). Most importantly, SUMO1 can be covalently attached to other proteins by a mechanism that resembles ubiquitination. The SUMO1 targets interact with hydrophobic and aromatic amino acids in SUMO1, which includes Phenylalanine 36, Valine 38 and Leucine 47, located in the groove of the SUMO1 molecule between the α-helix and β2 strand (Baba, et al., 2005).

Human SUMO1 sub-network in the Estrella database contains 80 interactors, out of 47 known structure interactors in the PDB, (41 non-redundant), and 7 have a

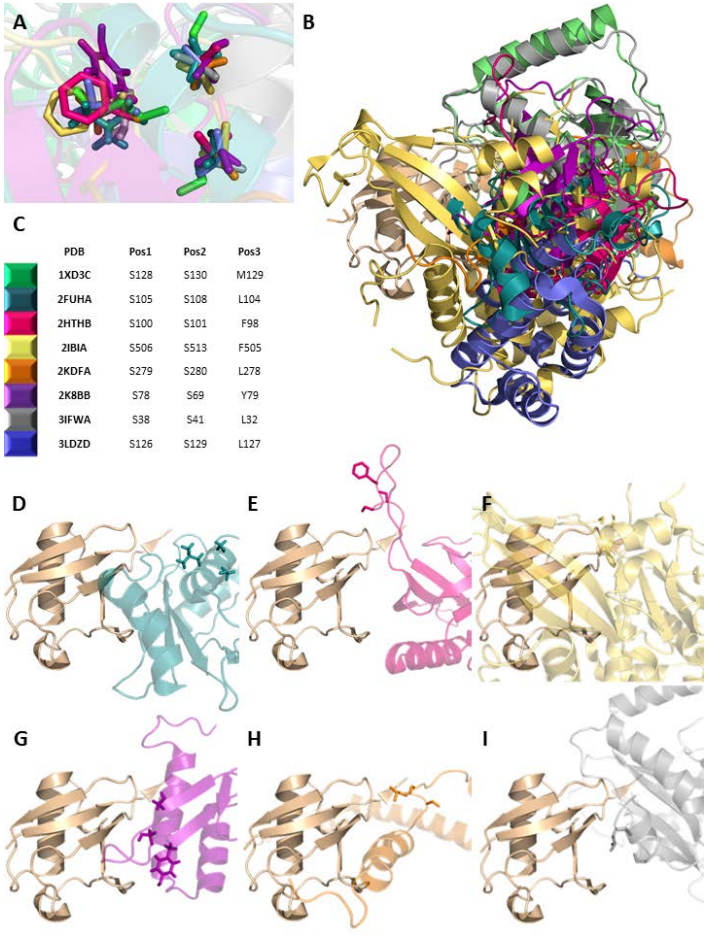


Figure 7: Ubiquitin local structure similarities found in Cluster 1**.** (A) Superposition of the similar residues, listed in (C). (B) Superposition of complexes using as reference ubiquitin, showing a shared binding site. Figures D-H depicts the local similarities found in cluster 1 among ubiquitin interactors

in solid sticks. Always ubiquitin protein is shown in solid cartoon wheat color, while (D) 2FUHA is in deep teal; (E) 2HTHB in hot pink (F) 2IBIA in yellow; (G) 2K8BB in purple; (H) 2KDFA in orange; (I) 3IFWA in grey.

experimentally determined structures including the interactor and the central protein, SUMO1 in this case (**Table 7**).

PDB Code	Ubiquitin Chain/s	Interactor Chain/s	UniProt Entry Name/s	Interactor Name
2IY1	2IY1B|D	2IY1A|C	SENP1_HUMAN	Sentrin-specific protease 1
2IO2	2IO2B	2IO2C	RGP1_HUMAN	Retrograde Golgi transport protein RGP1 homolog
1Z5S	1Z5SB	1Z5SA	UBE2I_HUMAN	Sumo-conjugating enzyme UBC9
1Z5S	1Z5SB	1Z5SD	RBP2_HUMAN	E3 SUMO-protein ligase RanBP2
3KYC	3KYCD	3KYCB	SAE2_HUMAN	Sumo-activating enzyme subunit 2
2KQS	2KQSA	2KQSB	DAXX_HUMAN	Death domain-associated protein 6
2ASQ	2ASQA	2ASQB	PIAS2_HUMAN	E3 SUMO-protein ligase PIAS2

Table 7: Experimentally determined complexes containing SUMO1. By column: PDB entry, SUMO1 chain/s, interactors chain/s, UniProt accession number, interactors UniProt entry name and protein name.

Superposing the SUMO1 of all the complexes, I identified an overlapping binding site comprising the 1Z5SD, 3KYCB, 2KQSB, 2ASQA protein chains and interacting with SUMO1 in the groove formed between the second β-strand and the first α-helix. Upon submitting these interactors to Estrella, I identified them as mutually exclusive and I found the best local structural similarity in cluster 1; this involves an overlapping binding interface (**Figure 8**), with Isoleucine (I) in position 1, Leucine (L) in position 2 and Aspartate (D) in position 3.

Case study CDC42

CDC42 belongs to the Rho subfamily of the Ras superfamily of GTPases that act as molecular switches in the control of a variety of eukaryotic processes (Hall, 1990). The major functions of CDC42 seem to be in regulating the rearrangements of the actin cytoskeleton in response to extracellular and intracellular signals as well as in modulating protein kinase cascades that result in the transcriptional activation of genes required for growth control.

In the Estrella database, CDC42 sub-network in human contains 78 interactors, 43 with a known structure in the PDB, 35 non-redundant, and 6 are

experimentally determined structures of complexes including the CDC42 protein (**Table 8**).

By superposing all complexes found using CDC42 as reference, I distinguished two different binding interfaces. The first, coinciding with CDC42 SWITCH I or effector domain and covering from the residue 26 to 50, interacts with three proteins: Serine/threonine-protein kinase PAK 6 (Q9NQU5), Partitioning defective 6 homolog beta (Q9JK83) and Activated CDC42 kinase 1 (Q07912), which respectively correspond with the following PDB entries: 2OBDB, 1NF3C and 1CF4B. Partitioning defective 6 homolog beta (Q9JK83) is a mouse protein that shares a 91% identity and 94% similarity and covers 100% of its homologous in human (Q9BYG5). Consequently, I included this structure due to the highly resemblance of the two proteins.

When submitting the SWITCH I domain interactors to Estrella, I obtained 100% of the clusters grouping the three proteins together as mutually exclusive interactors. The best local structure similarity was found in cluster 1 (**Figure 9**) comprising the following amino acids: Isoleucine (I) in position 1, Serine (S) in position 2, and Proline (P) in position 3.

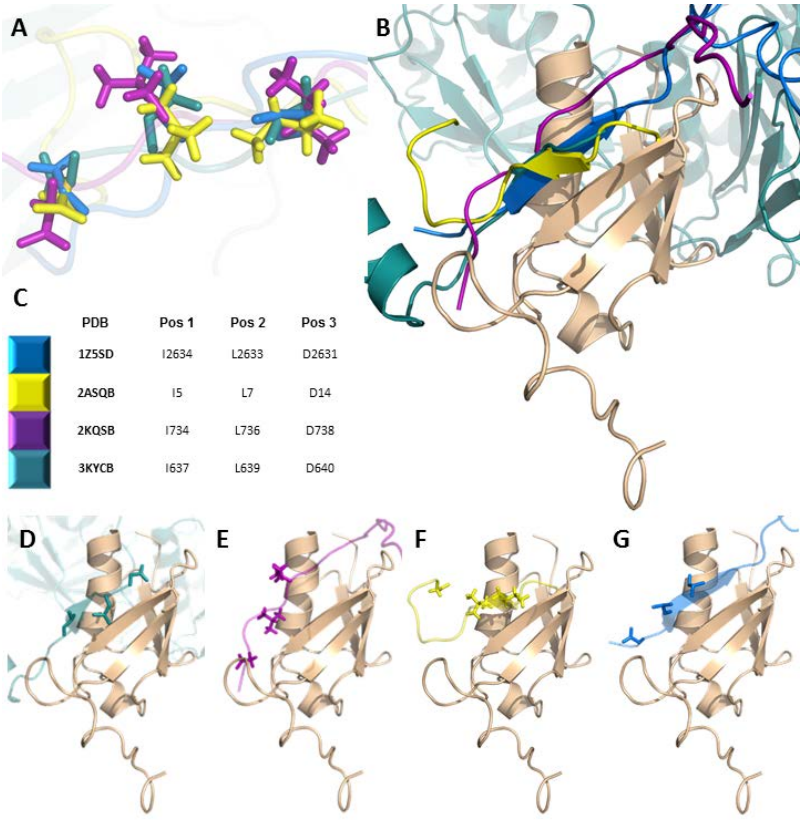


Figure 8: SUMO1 local structure similarities found in cluster 1. (A) Superposition of the similar residues, listed in (C). (B) Superposition of PDB complexes using as reference SUMO1, showing a shared binding site. Figures D-G depicts in solid sticks the local similarities found in cluster 1 among SUMO1 interactors. Always SUMO1 protein is shown in solid cartoon wheat color, meanwhile (D) 3KYCB in cyan; (E) 2KQSB in purple (F) 2ASQB in yellow; (G) 1Z5SD in blue.

PDB Code	Ubiquitin Chain/s	Interactor Chain/s	UniProt Entry Name/s	Interactor Name
1GRN	1GRNA	1GRNB	RHG01_HUMAN	Rho GTPase-activating protein 1
2ODB	2ODBA	2ODBB	PAK6_HUMAN	Serine/threonine-protein kinase PAK 6
1CEE	1CEEA	1CEEB	WASP_HUMAN	Wiskott-Aldrich syndrome protein
2WM9	2WM9B	2WM9A	DOCK9_HUMAN	Dedicator of cytokinesis protein 9
1CF4	1CF4A	1CF4B	ACK1_HUMAN	Activated CDC42 kinase 1
2DFK	2DFKB	2DFKA	ARHG9_RAT	Rho guanine nucleotide exchange factor 9

Table 8: Experimentally determine complexes containing CDC42. By column: PDB entry, CDC42 chain/s, interactors chain/s, UniProt accession number, interactors UniProt entry name and protein name.

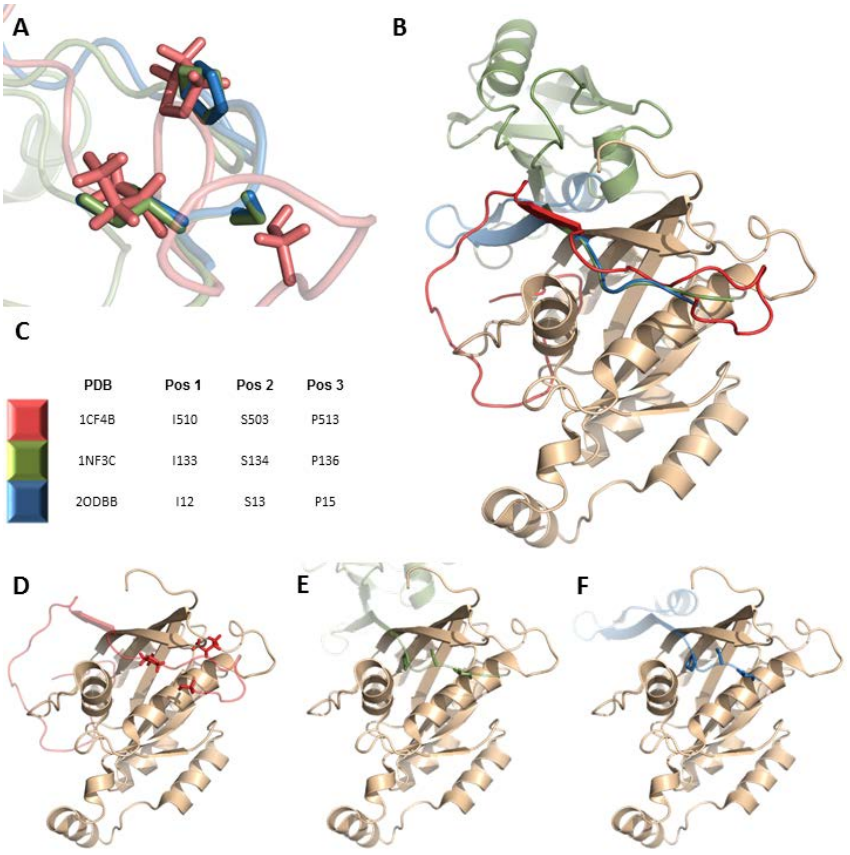


Figure 9: CDC42 local structure similarities found in Cluster 1. (A) Superposition of the similar residues, listed in (C). (B) Superposition of complexes using as reference Cdc42, showing a shared binding site. Figures D-F depicts in solid sticks the local similarities found in Cluster 1 among CDC42 interactors. Always CDC42 protein is shown in solid cartoon wheat color, meanwhile (D) 1CF4B in red; (E) 1NF3C in green; (F) 2ODBB in blue.

Discussion

It is becoming clear that combining the results of high throughput experiments and of computational analysis is a powerful strategy for transforming the ever-growing amount of information that we are accumulating into knowledge.

In this thesis I have described the application and the results of a straightforward idea: if two or more proteins interact with the same central protein, they might do so using the same central protein interface, in which case they might share similarity in their binding region. If such cases can be detected, it can be concluded that their interaction has to be mutually exclusive (i.e. not simultaneous) and that the region identified by similarity is indeed the interaction site.

I have tested the idea using seven different interactomes from different organisms. The data are stored in a publicly available database, which I hope will be useful to life scientists. The method provides very satisfactorily results, especially since it has a rather high specificity (above 85%), thereby ensuring that scientists

interested in a given biological process can retrieve essentially all of the *bona fide* mutually exclusive interactions in order to analyze them. Equally important is, in my point of view, that only in 13% of the cases the method incorrectly identifies more than one protein as part of a mutually exclusive interaction in a sub-network, and this implies that the number of necessary validating experiments is reduced.

Another observation that can be made from the results presented here is that the coverage of experimentally determined structures starts to be sufficient to allow their use in combination with different types of high throughput experiments.

Finally, the ever growing number of experimentally determined structures and of protein-protein interaction experiments, combined with the strategy presented here, also implemented in a completely automatic fashion and publicly accessible, is likely to add significant value to data produced in high-throughput experiments.

Bibliography

Alfarano, C. et al., 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res,* 1 Jan, 33(Database issue), pp. 418-24.

Aloy, P. & Russell, R., 2006. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol,* Mar, 7(3), pp. 188-97.

Armon, A., Graur, D. & Ben-Tal, N., 2001. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol,* 16 Mar, 307(1), pp. 447-63.

Arnold, E. & Rossmann, M., 1988. The use of molecular-replacement phases for the refinement of the human rhinovirus 14 structure. *Acta Crystallogr A,* 1 May, 44(3), pp. 270-82.

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* May, 25(1), pp. 25-9.

Ausiello, G. et al., 2008. FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinformatics,* 26 Mar, 9(Suppl 2), p. S2.

Ausiello, G., Via, A. & Helmer-Citterich, M., 2005. Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics*, 1 Dec, 6(Suppl 4), p. S5.

Baba, D. et al., 2005. Crystal structure of thymine DNA glycosylase conjugated to SUMO-1. *Nature*, 16 Jun, 435(7044), pp. 979-82.

Bader, G., Betel, D. & Hogue, C., 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 1 Jan, 31(1), pp. 248-50.

Bahadur, R., Chakrabarti, P., Rodier, F. & Janin, J., 2003. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 15 ov, 53(3), pp. 708-19.

Berggård, T., Linse, S. & James, P., 2007. Methods for the detection and analysis of protein-protein interactions. *Proteomics*, Aug, 7(16), pp. 2833-42.

Berman, H., 2008. The Protein Data Bank: a historical perspective. *Acta Crystallogr A*, Jan, 64(1), pp. 88-95.

Bernstein, F. et al., 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, 25 May, 112(3), pp. 535-42.

Boehr, D. & Wright, P., 2008. Biochemistry. How do proteins interact?. *Science*, 13 Jun, 320(5882), pp. 1429-30.

Brent, R. & Ptashne, M., 1985. A eukaryotic transcriptional activato rbearing the DNA specificity of a prokaryotic repressor. *Cell*, Dec, 43(3 Pt 2), pp. 729-36.

Brown, K. & Jurisica, I., 2005. Online predicted human interaction database. *Bioinformatics*, 1 May, 21(9), pp. 2076-82.

Chakrabarti, P. & Janin, J., 2002. Dissecting protein-protein recognition sites. *Proteins*, 15 May, 47(3), pp. 334-43.

Chatr-aryamontri, A. et al., 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Res*, Jan, 35(Database issue), pp. 572-4.

Chen, Y. & Xu, D., 2003. Computational analyses of high-throughput protein-protein interaction data. *Curr Protein Pept Sci*, Jun, 4(3), pp. 159-81.

Chen, Y. & Xu, D., 2005. Bioinformatics analysis for interactive proteomics. *Curr Protoc Protein Sci*, Dec.p. Chapter 25: Unit 25.1..

Chien, C., Bartel, P., Sternglanz, R. & Fields, S., 1991. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci USA*, 1 Nov, 88(21), pp. 9578-82.

Chothia, C., 1976. The nature of the accessible and buried surfaces in proteins. *J Mol Biol*, 25 Jul, 105(1), pp. 1-12.

Codd, E., 1998. A relational model of data for large shared data banks. 1970.. *MD Comput*, May-Jun, 15(3), pp. 162-6.

Cole, C. & Warwicker, J., 2002. Side-chain conformational entropy at protein-protein interfaces. *Protein Sci*, Dec, 11(12), pp. 2860-70.

Dayhoff, M., Schwartz, R. & Orcutt, B., 1978. A model for evolutionary change, in Atlas of Protein Sequence and Structure. Volume 5, pp. 345-58.

de Lichtenberg, U., Jensen, L., Brunak, S. & Bork, P., 2005. Dynamic complex formation during the yeast cell cycle. *Science*, 4 Feb, 307(5710), pp. 724-7.

de Vries, S. & Bonvin, A., 2008. How proteins get in touch: interface prediction in the study of biomolecular complexes. *Curr Protein Pept Sci*, Aug, 9(4), pp. 394-406.

Dikic, I., Wakatsuki, S. & Walters, K., 2009. Ubiquitin-binding domains - from structures to functions. *Nat Rev Mol Cell Biol*, Oct, 10(10), pp. 659-71.

Dundas, J. et al., 2006. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues.. *Nucleic Acids Res*, 1 Jul, 34(Web Server issue), pp. 116-8.

Ezkurdia, I. et al., 2009. Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform*, May, 10(3), pp. 233-46.

Fields, S. & Song, O., 1989. A novel genetic system to detect protein-protein interactions. *Nature*, 20 Jul, 340(6230), pp. 245-6.

Fields, S. & Sternglanz, R., 1994. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet*, Aug, 10(8), pp. 286-92.

Free, R., Hazelwood, L. & Sibley, D., 2009. Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy. *Curr Protoc Neurosci*, Jan.p. Chapter 5:Unit 5.28.

Fromont-Racine, M., Rain, J. & Legrain, P., 1997. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat Genet*, Jul, 16(3), pp. 277-82.

Ge, H., Liu, Z., Church, G. & Vidal, M., 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet*, Dec, 29(4), pp. 482-6.

Giot, L. et al., 2003. A protein interaction map of Drosophila melanogaster. *Science*, 5 Dec, 302(5651), pp. 1727-36.

Glaser, F., Steinberg, D., Vakser, I. & Ben-Tal, N., 2001. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 1 May, 43(2), pp. 89-102.

Gonzalez-Begne, M. et al., 2009. Proteomic analysis of human parotid gland exosomes by multidimensional protein identification technology (MudPIT). *J Proteome Res,* Mar, 8(3), pp. 1304-14.

Goodsell, D. & Olson, A., 2000. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct,* Jun, Volume 29, pp. 105-153.

Grote, J., Dankbar, N., Gedig, E. & Koenig, S., 2005. Surface plasmon resonance/mass spectrometry interface. *Anal Chem,* 15 Feb, 77(4), pp. 1157-62.

Güldener, U. et al., 2006. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res,* 1 Jan, 34(Database issue), pp. 436-41.

Hall, A., 1990. The cellular functions of small GTP-binding proteins. *Science,* 10 Aug, 249(4969), pp. 635-40.

Han, J. et al., 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature,* 1 Jul, 430(6995), pp. 88-93.

Hart, S. & Riba-Garcia, I., 2004. The analysis of tandem mass spectrometric datasets: high-throughput investigations require high-quality validation. *Drug Discov Today,* 1 May, 9(9), pp. 391-2.

Henikoff, S. & Henikoff, J., 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA,* 15 Nov, 89(22), pp. 10915-9.

Hermjakob, H. et al., 2004. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 1 Jan, 32(Database issue), pp. 452-5.

Hershko, A. & Ciechanover, A., 1982. Mechanisms of intracellular protein breakdown. *Annu Rev Biochem,* Volume 51, pp. 335-64.

Hobohm, U. & Sander, C., 1994. Enlarged representative set of protein structures. *Protein Sci,* Mar, 3(3), pp. 522-4.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C., 1992. Selection of representative protein data sets. *Protein Sci,* Mar, Issue 1, pp. 409-17.

Hope, I. & Struhl, K., 1986. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. *Cell,* 12 Sep, 46(6), pp. 885-94.

Hubbard, S. & Thornton, J., 1993. NACCESS. *Department of Biochemistry and Molecular Biology. University College London.*

Humphery-Smith, I., Cordwell, S. & Blackstock, W., 1997. Proteome research: complementarity and limitations with respect to the RNA and DNA worlds. *Electrophoresis,* Aug, 18(8), pp. 1217-42.

Ivanov, A., Zgoda, V. & Archakov, A., 2010. Technologies of Protein Interactomics: A Review. *Russian Journal of Bioorganic Chemistry,* 4 Aug, 37(1), pp. 8-21.

Jain, E. et al., 2009. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics,* 8 May.10(136).

Janin, J. & Chothia, C., 1990. The structure of protein-protein recognition sites. *J Mol Chem,* 25 Sep, 265(27), pp. 16027-30.

Janin, J., Miller, S. & Chothia, C., 1988. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol,* 5 Nov, 204(1), pp. 155-64.

Jeong, H., Mason, S., Barabási, A. & Oltvai, Z., 2001. Lethality and centrality in protein networks. *Nature,* 3 May, 411(6833), pp. 41-2.

Jones, S. & Thornton, J., 1996. Principles of protein-protein interactions. *Proc Natl Acad Sci USA,* 9 Jan, 93(1), pp. 13-20.

Jones, S. & Thornton, J., 1997. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol,* 12 Sep, 272(1), pp. 121-32.

Karlsson, R., 2004. SPR for molecular interaction analysis: a review of emerging application areas. *J Mol Recognit,* May-Jun, 17(3), pp. 151-61.

Keegan, L., Gill, G. & Ptashne, M., 1986. Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science,* 14 Feb, 231(4739), pp. 699-704.

Kelley, R. & Ideker, T., 2005. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol*, May, 23(5), pp. 561-6.

Kerrien, S. et al., 2007. IntAct--open source resource for molecular interaction data. *Nucleic Acids Res*, Jan, 35(Database issue), pp. 561-5.

Keskin, O., Gursoy, A., Ma, B. & Nussinov, R., 2008. Principles of protein-protein interactions: what are the preferred ways for proteins to interact?. *Chem Rev*, Mar, 108(4), pp. 1225-44.

Kim, P., Lu, L., Xia, Y. & Gerstein, M., 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 22 Dec, 314(5807), pp. 1938-41.

Koshland, D., 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci USA*, Feb, 44(2), pp. 98-104.

LaCount, D. et al., 2005. A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature*, 3 Nov, 438(7064), pp. 103-7.

Larsen, T., Olson, A. & Goodsell, D., 1998. Morphology of protein-protein interfaces. *Structure*, 15 Apr, 6(4), pp. 421-7.

Laskowski, R., 1995. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph*, Oct, 13(5), pp. 323-30, 307-8.

Laskowski, R. et al., 1997. PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem Sci*, Dec, 22(12), pp. 488-90.

Lee, B. & Richards, F., 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, 14 Feb, 55(3), pp. 379-400.

Lee, I., Date, S., Adai, A. & Marcotte, E., 2004. A probabilistic functional network of yeast genes. *Science*, 28 Nov, 306(5701), pp. 1555-8.

Levy, E., Pereira-Leal, J., Chothia, C. & Teichmann, S., 2006. 3D complex: a structural classification of protein complexes. *PLoS Comput Biol*, 17 Nov, 2(11), p. e155.

Liang, S., Zhang, C., Liu, S. & Zhou, Y., 2006. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res*, 7 Aug, 34(13), pp. 3698-707.

Li, S. et al., 2004. A map of the interactome network of the metazoan C. elegans. *Science*, 23 Jan, 303(5657), pp. 540-3.

Lo Conte, L., Chothia, C. & Janin, J., 1999. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 5 Feb, 285(5), pp. 2177-98.

Ma, B., Kumar, S., Tsai, C. & Nussinov, R., 1999. Folding funnels and binding mechanisms. *Protein Eng*, Sep, 12(9), pp. 713-20.

Masters, S., 2004. Co-immunoprecipitation from transfected. *Methods Mol Biol*, Volume 261, pp. 337-50.

Mayer, R., Landon, M. & Layfield, R., 1998. Ubiquitin superfolds: intrinsic and attachable regulators of cellular activities?. *Fold Des*, 3(5), pp. R97-9.

Mihel, J. et al., 2008. PSAIA - protein structure and interaction analyzer. *BMC Struct Biol*, 9 Apr.8(21).

Mishra, G. et al., 2006. Human protein reference database--2006 update. *Nucleic Acids Res*, Jan, 34(Database issue), pp. 411-4.

Monod, J., 1965. Reflections on the relationship between the structure and function of globular proteins. *Annee Biol*, Mar-Apr, Volume 59, pp. 231-40.

Moore, P., 2001. The ribosome at atomic resolution. *Biochemistry*, 20 Mar, 40(11), pp. 3243-50.

Murzin, A., Brenner, S., Hubbard, T. & Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 7 Apr, 247(4), pp. 536-40.

Neuvirth, H., Raz, R. & Schreiber, G., 2004. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*, 16 Apr, 338(1), pp. 181-99.

Nooren, I. & Thornton, J., 2003. Diversity of protein-protein interactions. *EMBO*, 15 Jul, 22(14), pp. 3486-92.

Orchard, S. et al., 2007. Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics*, Sep, Suppl(1), pp. 28-34.

Orengo, C. et al., 1997. CATH--a hierarchic classification of protein domain structures. *Structure*, 15 Aug, 5(8), pp. 1093-108.

Pagel, P. et al., 2005. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, Mar, 21(6), pp. 832-4.

Peri, S. et al., 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, Oct, 13(10), pp. 2363-71.

Petroski, M., 2008. The ubiquitin system, disease, and drug discovery. *BMC Biochem*, 21 Oct, 9(Suppl 1), p. S7.

Phizicky, E. & Fields, S., 1995. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, Mar, 59(1), pp. 94-123.

Pintar, A., Carugo, O. & Pongor, S., 2002. CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics*, Jul, 18(7), pp. 980-4.

Pintar, A., Carugo, O. & Pongor, S., 2003. DPX: for the analysis of the protein core. *Bioinformatics*, 22 Jan, 19(2), pp. 313-4.

Razick, S., Magklaras, G. & Donaldson, I., 2008. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, Sep, 30(9), p. 405.

Rich, R. & Myszka, D., 2000. Advances in surface plasmon resonance biosensor analysis. *Curr Opin Biotechnol*, Feb, 11(1), pp. 54-61.

Rual, J. et al., 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 20 Oct, 437(7062), pp. 1173-8.

Ruepp, A. et al., 2008. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, Jan, 36(Database issue), pp. 646-50.

Salwinski, et al., 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 1 Jan, 32(Database issue), pp. 449-51.

Stark, C. et al., 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 1 Jan, 34(Database issue), pp. 535-9.

Stelzl, U. et al., 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 23 Sep, 122(6), pp. 957-68.

The UniProt Consortium, 2011. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*, Jan, 39(Database issue), pp. 214-9.

Tong, A. & Boone, C., 2006. Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol*, Volume 313, pp. 171-92.

Tong, A. et al., 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 14 Dec, 294(5550), pp. 2364-8.

Tong, A. et al., 2004. Global mapping of the yeast genetic interaction network. *Science*, 6 Feb, 303(5659), pp. 808-13.

Tuncbag, N. et al., 2009. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform*, May, 10(3), pp. 217-32.

Uetz, P. et al., 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature,* 10 Feb, 403(6770), pp. 623-7.

Valdar, W. & Thornton, J., 2001. Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol,* 19 Oct, 313(2), pp. 399-416.

von Eichborn, J., Bourne, P. & Preissner, R., 2011. Cobweb: a Java applet for network exploration and visualisation. *Bioinformatics,* 15 Jun, 27(12), pp. 1725-6.

von Mering, C. et al., 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature,* 23 May, 417(6887), pp. 399-403.

Wang, G. & Dunbrack, R. J., 2003. PISCES: a protein sequence culling server. *Bioinformatics,* 12 Aug, 19(12), pp. 1589-91.

Yaciuk, P., 2007. Co-immunoprecipitation of protein complexes. *Methods Mol Med,* Volume 131, pp. 103-11.

Yan, Y. & Marriott, G., 2003. Analysis of protein interactions using fluorescence technologies. *Curr Opin Chem Biol,* Oct, 7(5), pp. 635-40.

Zemla, A., 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res,* 1 Jul, 31(13), pp. 3370-4.

Zemla, A., Venclovas, C., Moult, J. & Fidelis, K., 1999. Processing and Analysis of CASP3 Protein Structure Predictions. *Proteins,* Suppl(3), pp. 22-9.

Zgoda, V. et al., 2009. Proteomics of mouse liver microsomes: performance of different protein separation workflows for LC-MS/MS. *Proteomics,* Aug, 9(16), pp. 4102-5.

Zhou, H. & Qin, S., 2007. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics,* 1 Sep, 23(17), pp. 2203-9.

Zhou, H. & Shan, Y., 2001. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins,* 15 Aug, 44(3), pp. 336-43.

Zhukov, A. et al., 2004. Integration of surface plasmon resonance with mass spectrometry: automated ligand fishing and sample preparation for MALDI MS using a Biacore 3000 biosensor. *J Biomol Tech,* Jun, 15(2), pp. 112-9.