On the study of 3D structure of proteins for developing new algorithms to complete the interactome and cell signalling networks

Joan Planas Iglesias

TESI DOCTORAL UPF / 2012

DIRECTOR DE LA TESI

Dr. Baldomero Oliva Miguel

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

UNIVERSITAT POMPEU FABRA

Al meu pare,

que una tarda de primavera va salpar cap un lloc ben llunyà.

Acknowledgements

Firstly and foremost I need to thank my family, some wonderful and discreet people that always held me up and encouraged me to reach my goals. Although discretion should be always rewarded with the same coin, I must need to name hear Father and Montse, to whom I wont be able to express my thankfulness anymore if not by these words. Only the will of all my relatives has made this work possible.

To my friends, Mireya and the countless rest of them. It is said that devil is in the little things, and ordinality is one of such. So I won't go through an unending enumeration of my friends' names. Anyhow, all of them know that I needed their warmth -laughs, smiles, and counsel, to reach this far. And they also know that I have many good reasons to keep them deep in my heart.

This thesis wouldn't be the same if not for the whole GRIB troupe and my fellow lab-mates. I have no room to type the names of all people from GRIB, but their aid and assistance have been priceless. Nuria, Ramón, Josep, Carme, David, Armando, Daniel, Jaume, Javi, David, Oriol, Aggeliki, Valerio, Emre, Elisenda, Marc, Manu, Jascha, Daniel, Alessandra, Roger, Attila, Billur, and Bernat have shared much more than an office with me. All their names well deserve to be mentioned here; their help and advice has always been most welcomed.

I have to acknowledge my collaborators from the University of Edinburgh, Prof. Peter Ghazal and Dr. Kevin Robertson. Their contribution to this work has been invaluable.

My last words of acknowledgment (truly in a place of honour) are for my PhD. thesis supervisor, Prof. Baldo Oliva, who is at least as much responsible of this work as I am. I have many, many things to thank him, but here I will only emphasize that during this troubled and long-lasting period, he has given me all the reasons to consider that research is just a wonderful adventure. Thanks, Baldo!

Once again, my warmest thanks to all of you.

Abstract

Proteins are indispensable players in virtually all biological events. The functions of proteins are determined by their three dimensional (3D) structure and coordinated through intricate networks of protein-protein interactions (PPIs). Hence, a deep comprehension of such networks turns out to be crucial for understanding the cellular biology. Computational approaches have become critical tools for analysing PPI networks. *In silico* methods take advantage of the existing PPI knowledge to both predict new interactions and predict the function of proteins. Regarding the task of predicting PPIs, several methods have been already developed. However, recent findings demonstrate that such methods could take advantage of the knowledge on non-interacting protein pairs (NIPs). On the task of predicting the function of proteins,the Guilt-by-Association (GBA) principle can be exploited to extend the functional annotation of proteins over PPI networks. In this thesis, a new algorithm for PPI prediction and a protocol to complete cell signalling networks are presented. iLoops is a method that uses NIP data and structural information of proteins to predict the binding fate of protein pairs. A novel protocol for completing signalling networks –a task related to predicting the function of a protein, has also been developed. The protocol is based on the application of GBA principle in PPI networks.

Resum

Les proteïnes tenen un paper indispensable en virtualment qualsevol procés bioloògic. Les funcions de les proteïnes estan determinades per la seva estructura tridimensional (3D) i són coordinades per mitjà d'una complexa xarxa d'interaccions protiques (en anglès, *protein-protein interactions*, PPIs). Així doncs, una comprensió en profunditat d'aquestes xarxes és fonamental per entendre la biologia cel·lular. Per a l'anàlisi de les xarxes d'interacció de proteïnes, l'ús de tècniques computacionals ha esdevingut fonamental als darrers temps. Els mètodes *in silico* aprofiten el coneixement actual sobre les interaccions proteiques per fer prediccions de noves interaccions o de les funcions de les proteïnes. Actualment existeixen diferents mètodes per a la predicció de noves interaccions de proteines. De tota manera, resultats recents demostren que aquests mètodes poden beneficiar-se del coneixement sobre parelles de proteïnes no interaccionants (en anglès, *non-interacting pairs*, NIPs). Per a la tasca de predir la funció de les proteïnes, el principi de "culpable per associació" (en anglès, *guilt by association*, GBA) és usat per extendre l'anotació de proteïnes de funció coneguda a través de xarxes d'interacció de proteïnes. En aquesta tesi es presenta un nou mètode pre a la predicció d'interaccions proteiques i un nou protocol basat per a completar xarxes de senyalització cel·lular. iLoops és un mètode que utilitza dades de parells no interaccionants i coneixement de l'estructura 3D de les proteïnes per a predir interaccions de proteïnes. També s'ha desenvolupat un nou protocol per a completar xarxes de senyalització cel·lular, una tasca relacionada amb la predicció de les funcions de les proteïnes. Aquest protocol es basa en aplicar el principi GBA a xarxes d'interaccions proteiques.

Preface

In 2001 the first draft of the human genome was disclosed, making the "book of life" available to the whole scientific community. Since then, lots of efforts have been done to understand how this vast succession of nucleotides translated into observable phenomena in living beings. However, very little of that book is currently understood, because of the countless layers of regulation it has. And despite of this, all what is needed to understand the genome is enclosed in it…

In this regard, the regulatory role of non-coding nucleic acids has recently been unravelled. However, proteins play a prominent role in regulating the expression from genes. For instance, the positioning of nucleosomes determines which genomic regions can be read. It has bee shown that particular combinations of proteins enhance and silence the transcription of certain gene exons. Uppermost, transcription factors are the proteins that regulate the activation of genes.

To perform all these functions proteins need to interact with others, either to form complexes or to recognize precise targets of their action. For instance, a particular transcription factor may activate one gene or other depending on which protein interactions it performs. Other cellular processes strongly rely in the formation of protein-protein interactions as well. The recognition and binding of particular elements in signalling pathways is crucial for the adaptation of the cell to its environment. Also, formidable protein complexes are assembled to constitute the basic cellular machinery.

Since protein interactions are crucial for the cell survival, the ability of the proteins to interact with others and the particular partners they interact with are among the most important characteristics of a protein. In this context, the analysis and prediction of protein–protein interactions become central topics to achieve a better understanding of the cell and living organisms.

Table of contents

List of figures

List of tables

1 Introduction

1.1 Networks and biological networks

Live beings store the information required for developing themselves in form of deoxyribonucleic acid (DNA), a book of instructions that is carefully preserved in the nucleus of each of our cells. The exact duplication of this information from generation to generation assures the genetic continuity of species. The DNA is a sequence of nucleotides that are arranged in genes, the hereditary units that devise all identifiable traits of an organism. However, the main effectors of biological activities in the cell are not the genes but their products: different forms of ribonucleic acid (RNA) and proteins. By means of two differentiated processes, transcription and translation, the information encoded in genes is successively conveyed to RNA and proteins. During transcription, DNA is copied into messenger RNA (mRNA), which carries the instructions form DNA specifying the order of amino acids for protein biosynthesis. Then, starting the translation, the mRNA is read by the ribosome, a heterogeneous complex formed by several proteins and ribosomal RNA (rRNA). To read the mRNA and produce a protein, the ribosome recruits a third type of RNA molecule, the transference RNA (tRNA). This molecule is able to bind a particular mRNA triplet and to couple an amino acid specific to that triplet. To produce a protein, the ribosome "reads" the mRNA, and for each of its triplets it recruits a different tRNA, which in turn carries the particular amino acid encoded by the mRNA triplet. Hence, while reading the mRNA string, different amino acids are brought to the ribosome and become linked by means of peptide bonds, building the protein's sequence or primary structure.

The theoretical bases of this transfer of information were established decades ago (1,2), conforming the central dogma of biology[1]. Even though, transferring the information from the DNA to its products, RNAs and proteins, is not enough to make the cell work and organisms live (3). The process by which a gene is "turned on" to yield its specific product (RNA or protein) is referred as "gene expression". Different genes have their

[1] This idea was initially proposed by Francis Crick in 1956 in the letter "On protein synthesis". It was finally published in 1958 under the same name, and later revised in 1970.

expression heavily controlled by a process known as “gene regulation”, in order to generate their products in a precise temporal pattern and thus coordinate their tasks (4). Once produced, most proteins need to acquire a particular three-dimensional (3D) conformation through a process named folding to perform their functions (5,6). After folding, proteins can associate to the impressive cell machinery, responsible of crucial functions in the cell (7,8). Indeed, the formation of such protein-protein interactions (PPIs) is closely related to the regulation of several cellular functions (9). Furthermore, in order to adapt themselves to the surrounding environment, cells respond to external stimuli by means of signalling pathways, intricate connections that convey environmental information from the cellular membrane to its nucleus. These pathways involve both PPIs and post-translational modifications of the proteins, including phosphorilation, metylation, and acetylation (10,11). The signal is propagated until a particular protein type, a transcription factor, is activated, allowing its translocation to the cell nucleus where it can regulate or activate a specific genetic program. Finally, to perform all these functions the cell needs to produce nucleotides, amino acids, lipids and oligosaccharides, to obtain energy, to store it, and to manage its consumption; in other words, keep their homeostasis. Enzymes, proteins capable of converting specific compounds (their substrates) into others (their products) during an enzymatic reaction, are responsible of the accomplishment of these tasks. Enzymatic reactions are chained forming metabolic pathways, functional units aimed to accomplish the conversion of one initial substrate to one final product (12). For instance, acetate derived from carbohydrates, lipids, and proteins, is oxidized during the citric acid cycle producing carbon dioxide and energy. Although classical biochemistry depicted metabolic pathways as separated functional units (13,14), there exists a huge pool of smaller metabolic units and chemical compounds that can be used by numerous metabolic pathways (15). Hence, different metabolic pathways are indeed interconnected, forming a complex system known as metabolic network.

All these processes can be depicted as networks, where nodes represent biological entities (genes, amino acids, proteins, compounds), and edges the relationships established between them (regulation, contact, interaction, consumption). *In silico* tools are helpful instruments to study and analyze networks allowing the discovery of unravelled properties within them and the prediction of novel relationships between the

biological units represented in the networks. In summary, computational tools, which are used to study protein structure and biological networks, can also give new insights on how cells work and living organisms live. The scientific disciplines that study, develop and apply such tools are bioinformatics and computational biology.

1.1.1 Definitions of bioinformatics and computational biology

According to the USA National Institute of Health (NIH), bioinformatics and computational biology are defined (16) as follows:

- *Bioinformatics*: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organize, archive, analyze or visualize such data.
- *Computational Biology*: The development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, behavioural and social systems.

The same defining committee stressed similarities and differences between the disciplines (16):

"Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface."

Due to this overlap, the terms *bioinformatics* and *computational biology* will be used indistinctively along this thesis to refer to the development and application of computational tools (including theoretical methods and mathematical models) for studying biological systems and data.

One of the fields that take more advantage of computational biology techniques is the study of biological networks. The following section is devoted to describe the mathematical description of a network and its properties, which are key for the exploitation of *in silico* approaches in the study of biological networks.

1.1.2 Definitions of networks and their properties[2]

The mathematical object that expresses the relationships among a series of objects is known as *graph*[3]. A *simple graph* can be defined as $G=(V,E,I)$, where V and E are disjoint finite sets and I is an incidence relation such that every element of E is incident with exactly two distinct elements of V and no two elements of E are incident to the same pair of elements of V. In this context, V and E are called the *vertex set*[4] and the *edges set* of G respectively. Recalling the image of a biological network, the vertices would represent the elements in such network (genes, amino acids, proteins, compounds) and the edges the relationships established among them (regulation, interaction, etc.).

The edges and vertices of a graph can have specific values assigned, denoting their relative strength, importance, or any other property that permits to establish a ranking or classification among them. Such edges and vertices are known as *labelled* -or *weighted* if the labelling property is a countable measure. If vertices in a graph G are divided into two disjoint sets U and V (i.e. they are assigned one of two different labels) and every

---

[2]This section is intended to provide a rough introduction to the mathematical representation of networks and their properties. All definitions herein presented were extracted from 17. Gross, J. and Yellen, J. (1999) *Graph Theory and Its Applications*. CRC Press, Boca Raton.

[3] In mathematical context, the English word *graph* is polysemous: it can denote a graph function (i.e. a plot) or a collection of dots and lines connecting some (possibly empty) subset of them. In this section, the word *graph* will be used for the second meaning.

[4] The mathematical term for referring to a dot in a graph is *vertex* (plural *vertices*); however, in biology-related disciplines such as computational biology vertices are also referred as *nodes*. Here, both terms will be used indistinctively.

edge in G connect a vertex in U to one in V, G is said to be *bipartite*. In addition, edges may be *directed* if denote directionality; in other words, a directed edge establishes a connection from node u to node v, but not from v to u (is incident to v but not to u). Note that a vertex is incident to its surrounding edges, but an edge may be incident to some of its surrounding nodes depending on its directionality. A graph is considered to be a *directed graph* if all its edges are directed.

Two vertices in a graph are regarded as *adjacent* if they are incident to a common edge. The set of *neighbours*, $N_G(v)$, of a vertex v is the set of vertices which are adjacent to v. A *walk* is an alternating sequence of vertices and edges, with each edge being incident to the vertices immediately preceding and succeeding it in the sequence. Thus, a walk comprising directed edges is constrained by their directionality. The *length* l of a walk is the number of edges that it uses. For a non-closed walk, $l = n–1$, where n is the number of vertices visited (a vertex is counted each time it is visited). For a closed walk, $l = n$ (the start/end vertex is listed twice, but is not counted twice). A walk with no repeated edges is called *trail* and it is known as a *path* if it has no repeated vertices. The *distance* from u to v, written $d_G(u,v)$, is the minimum length of any path from u to v. A walk is *closed* if the initial vertex is also the terminal vertex; a closed trail containing at least one edge is known as *cycle*. A graph is *cyclic* if it contains any cycles, and *acyclic* otherwise.

A *subgraph* of a graph G is a graph whose vertex set is a subset of that of G, and whose adjacency relation is a subset of that of G restricted to this subset. A graph G is *connected* if, for every pair of vertices u and v, there exists a path from *u to v; G* is *disconnected* otherwise. A *component* of G is defined as a maximal connected subgraph of G. In other words, a connected subgraph H is a component of graph G if H is not a proper subgraph of any connected subgraph of G. Thus, the only component of a connected graph is the entire graph and, intuitively, the components of a non-connected graph are the "whole pieces" it comprises. A *complete graph* K_n of order n is a simple graph with n vertices in which every vertex is adjacent to every other. In a graph G, a *clique* is any subgraph H that is complete. A *k-clique* is a clique of order k. A *maximal clique* is a clique that is not a subset of any other clique. The *clique number* $\omega(G)$ of a graph G is the order of a largest clique in G.

Tables 1.1 and 1.2 summarize the principal metrics of nodes and graphs used in computational biology respectivelly.

Table 1.1. Principal metrics of a node *v* in a graph *G*.

Property	**Symbol**	**Definition**
Degree	$d_G(v)$	Number of adjacent nodes.
Betweenness	$g_G(v)$	Number of shortest paths between any pair of nodes (*u, w)* in *G* that pass through *v*.
Eccentricity	$e_G(v)$	Maximum value of distance to any other node *u* in *G*
Wienner index	$W_G(v)$	Sum of the distances to each other node *u* in *G*
Clustering coefficient	$C_G(v)$	Number of edges connecting any neighbour *u* of *v* over the total number of possible edges connecting any neighbour *u* (i.e. as if *G* was a complete graph).

Table 1.2. Principal metrics of a graph *G*.

Property	**Symbol**	**Definition**
Radius	*rad(G)*	Minimum value $e(v)$ for any node *v* in *G*
Diameter	*dim(G)*	Maximum value $e(v)$ for any node *v* in *G*
Average degree	*d(G)*	Arithmetic mean of the degrees of all nodes *v* in *G*
Connectivity	$\kappa(G)$	Minimum number of nodes *v* that need to be removed to disconnect *G*

1.1.3 Topology of biological networks

Due to the network-wise nature of most cellular processes, the study of networks has become a booming field in theoretical biology. Barabasi and Oltvai described the understanding of "the structure and the dynamics of the complex intracellular web of interactions that contribute to the structure and function of a living cell" as a "key challenge for biology in the twenty-first century" (18). Their work was crucial for the early development of *systems biology*, which "aims to map out, understand, and model in quantifiable terms the topological and dynamic properties of the biological networks" (18). Indeed, recent advances in biological data collection (high throughput data such as

DNA or RNA micro-chips, protein chips and yeast two-hybrid screens) and bioinformatics techniques allowed addressing such challenge. Particularly, the integration of different data sources such as protein sequence, gene expression and PPIs has prompted a quick development of the understanding of the cell.

Perhaps one of the most important findings the study of biological networks has revealed is that their architecture is not arbitrary. A random network comprises N nodes, which are pair-wise connected with a probability p (19). In such networks, the distribution of the degree of their nodes is normal (see Figure 1.1 A). However, biological networks do not follow such pattern; instead they approximate a scale-free topology, which is characterized by a power-law degree distribution. In such networks, the probability that a node has k links follows $P(k) \sim k^{-\gamma}$, where γ is the degree exponent and, the probability that a node is highly connected (i.e. is a *hub* node) is statistically more significant than in a random graph (20) (see Figure 1.1 B). The first observations of the scale-free topology in biological networks were done on metabolic networks, where most metabolic compounds participate only in one or two reactions, but a few of them (such as pyruvate or coenzyme A) participate in a large number of reactions, becoming metabolic hubs (21,22). The same topology has been determined for protein interaction networks (23,24) (although it is currently being questioned (25,26)), regulatory networks (27-29), and signalling pathways (30).

The analysis of high-throughput biological data has revealed a modular organisation of cellular functionality (31), defined as "separability of the design into units that perform independently, at least to a first approximation" (32). Furthermore, it has been shown that small sub-networks (also known as motifs) can be re-used as basic bricks to build larger ones (33,34). However, enforcing a power-law degree distribution in theoretical models (scale-free networks) is not enough to reproduce the observed modular nature of biological networks. Scale-free networks are characterized by the presence of a small number of highly connected nodes (hubs) that, in large networks, generate a single integrated web in which the existence of fully separated modules is apparently impossible. Analyzing the metabolic network, Ravasz *et al.* pioneered a solution to this dilemma: the hierarchical network model (35). In short, this model consists in iterative replicas of a small and highly interconnected cluster of nodes (module). First level

replicas are loosely connected among them and to the original cluster, and together form a new unit –a larger module, that can be in turn replicated in the same manner. This model (Figure 1.1 Ca) is coherent with both the scale-free property of networks (Figure 1.1 Cb), and the modularity of biological networks, denoted by a clustering coefficient dependant on the degree of the node (Figure 1.1 Cc). This model is compatible with other biological networks including protein-protein interaction and gene regulatory networks (18).

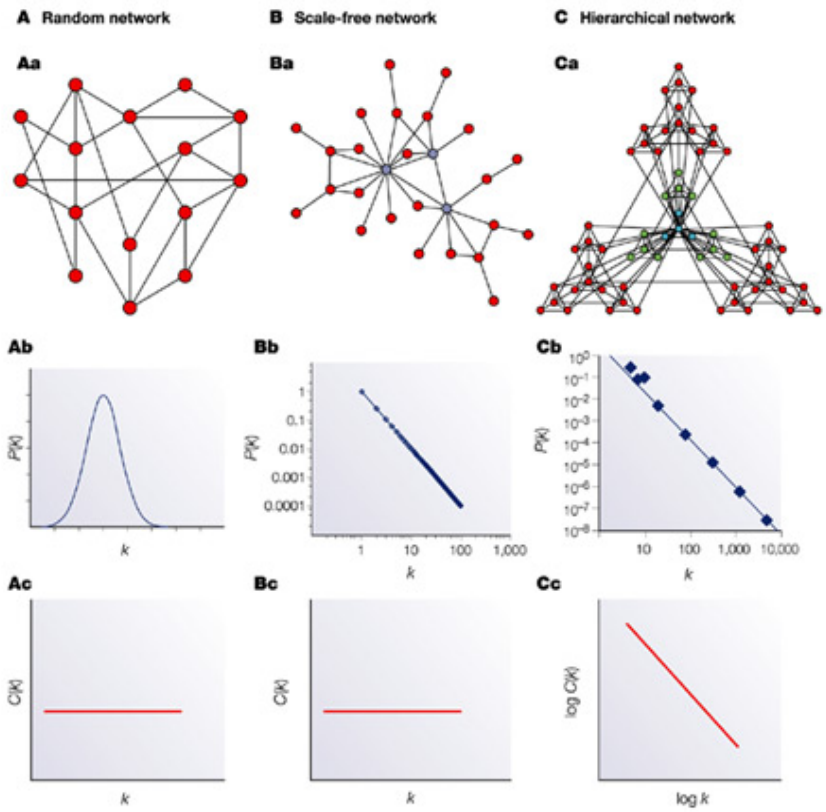


Figure 1.1. Network models in computational biology (obtained from (18)). Random, scale-free, and hierarchical networks are respectively represented in first, second and third column of the first row. Note that hub nodes are in light-shaded while non-hub nodes are dark-shaded. The second and third row represent the degree (*k*) distribution ad the clustering coefficient (*C(k)*)of their nodes.

1.1.4 Topology meets biology: uses of biological networks to extend functional annotation of genes and proteins.

Computational biology has taken advantage of the architecture of biological networks in several ways. A major breakthrough in the field of biological network analysis was the establishment of the *guilt-by-association* (GBA) concept. GBA principle suggests that genes and gene products with related functions tend to share properties such as genetic or physical interactions (36), and is indirect relationship with the modular property of biological networks. In other words, nodes that are in close proximity to others in a network are more likely to share functions. The basic idea of GBA was early exploited to predict the function of unnanotated genes in regulatory networks (37-39) and unnanotated proteins in protein-protein interaction networks (interactomes) (40,41). Furthermore, the GBA principle still underlies in current methods for predicting the function of genes (42) and proteins (43).

Regarding biological relevance of networks architecture, another significant milestone was reached when highly connected nodes (hubs) were related to cell survival in *Saccharomyces cerevisiae* (44). Hub proteins were quickly reported to be involved in human disease (45), and were also identified as key proteins in signalling pathways whose disregulation could prompt pathological conditions (46-48). Subsequent analyses revealed that the relationship between the degree of a node and its implication in pathology or cell death was actually more reflective of the number of distinct processes the node was involved in. For instance, Kim *et al.* showed that, in *Saccharomyces cerevisiae* interactome, highly connected proteins (i.e., hubs) with multiple binding interfaces were twice as likely to be essential as hubs with one or two interfaces (49).

The potentiality of GBA and the study of hub proteins to associate genes and proteins with disease is vast. Several methods have been developed to identify candidate disease genes based on the proximity to other known disease genes (seeds) in regulatory networks or interactomes. Such proximity can be defined by direct neighbourhood ((50,51)), shortest paths ((50,52)) or random walks ((50,53)) along the edges connecting a given seed with a disease candidate. The state of the art of this field has recently been reviewed in (54,55)), and associating genes and gene products to disease is still one of the main focuses of research in network biology (56,57).

1.1.5 Other applications of computational biology to biological networks

Other uses of biological networks in computational biology are specific to the particular type of network studied. This section covers the principal applications of computational biology to the most important biological networks in the cell:

i) gene regulatory networks,

ii) protein interaction networks (or interactomes),

iii) signalling networks,

iv) metabolic networks,

Reverse engineering allows the reconstruction of gene regulatory networks from time-series expression data (58). Several methods have been developed based on this idea, including probabilistic methods (mainly Bayesian networks (59)), correlation-based methods (i.e. WGCNA (60)), partial-correlation-based methods (i.e. SPACE (61) or GenNet (62)), and information-theory-based methods (i.e. ARACNE (63)). All of them have been recently reviewed in (64).

Besides functional annotation, interactomes have been applied to other objectives. Two inherent problems in the experimental detection of PPIs are incompleteness and noise (65) (see section 1.3.5). Bader et al. pioneered the study of protein interaction networks to estimate the reliability of interactions (66). Later on, Gavin et al. developed socio-affinity scores on de novo identified protein complexes to quantify the propensity of proteins to form partnerships (9). Novel and more generic approaches for scoring the reliability of PPs have been recently developed (67). Protein interaction networks can be used to predict novel protein interactions if combined with further information such as genomic context (68) or structural knowledge of proteins (69). Further more, insight on the interacting region can be gained with this approach (70,71). Even in absence of contextual information, the topology of interactomes can reveal putative interactions (72) or interacting regions of highly connected proteins in the network (73). Other applications of protein interaction networks include:

i) the identification of domain-domain interactions (74,75);

ii) the delineation of frequent interaction network motifs (76); and

iii) the comparison between mode organisms and humans (77).

Due to the fuzzy nature of cell signalling, the problem of completing signalling pathways has been classically addressed using PPI prediction approaches (78). However, most high-throughput experiments for protein interaction discovery result in little new knowledge regarding phosporilation events, a key process in cell signalling (79). This is because the transient nature of most PPIs in signalling pathways requires specific experimental approaches to pinpoint such interactions (80). The lack of high quality interaction data covering signalling proteins has been surpassed by the use of other functional association information (81). Nevertheless, recent advances in network biology have allowed approaching this problem from a network perspective (82).

Computational biology has played a major role in the reconstruction of metabolic networks. The combination of human manual curation of metabolic data and automated aimed to discover missing elements in metabolic pathways has allowed the compilation of comprehensive repositories of metabolic information (83,84). Computational applications of reconstructed metabolic networks have been recently reviewed in (85) and include:

i) contextualization of high-throughput data;

ii) guidance of metabolic engineering;

iii) directing hypothesis-driven discovery;

iv) interrogation of multi-species relationships, and

v) network property discovery,

The computational applications for different biological networks previously detailed rely in accurate descriptions of such networks. However, the very nature of biological data makes the task of obtaining such high-quality networks very difficult. The

following section is devoted to describe in detail this problem and the different approaches taken to solve it.

1.1.6 From genes to proteins: the problem of naming and identifying biological entities

Since first enunciation of the central dogma of biology (2), the paradigm for protein production has shifted from the "one gene, one protein" concept to a more complex view where from a single gene several products can be obtained. The roots of this spread of information can be found in mutations in the DNA sequence (including single point mutations, insertions, and deletions), splicing variants in the RNA transcripts and other post-translational mutations (PTMs) in the final protein product. It is a challenging task to integrate that high amount of variability in networks –a simple form for representing the complexity of cellular processes. However, the main difficulty in univocally representing biological entities in a network is that different interfaces for accessing biological data provide different identifiers for equivalent biological entities, a problem of which scientific community has been aware since long time ago (86).

In order to combine data from different sources and software applications, substantial effort has been spent. Regarding PPI data, the Human Proteome Organization (HUPO) (87) has developed PSI-MI (88) towards achieving standard formats to exchange data and well-defined protocols of PPIs. PSI-MI is an XML-based schema for the representation of molecular interactions. This schema represents PPIs and several associated attributes, such as their detection method, the role of each protein in the experiment, or the stoichiometry of the protein-partners of a complex. Other standards allow the gathering, storage and processing of other relationships among biomolecules. For instance, BioPax (89) focuses on biological pathways that include PPIs; and Systems Biology Markup Language (SBML) (90) is a XML based standard that represents computational models of any kind of biological network.

Despite the development of such standards, several biological data resources still use distinct identifiers for genes (or proteins) encumbering a non-redundant unification, i.e. univocally identifying each independent biological unit in the network. This challenge

has been tackled by different initiatives. Some of them have built new datasets by combining and eliminating redundancies from several resources, such as PINA (91), BIANA (92), APID (93) or iRefIndex (94) (figure 1.2 A). Furthermore, PSICQUIC (95) is the result of a recent effort to standardize the access to molecular interaction databases programmatically by using a query language system. Organizations such as EBI offer PSICQUIC access to their network repositories (96).

Another strategy to integrate data from different sources in a single network is to develop frameworks where the data is stored locally and the user the user sets the criteria for the unification rules. This is the case of BIANA (92) and ONDEX (97). For example, BIANA allows the users to choose the data and select the features (or identifiers) used for the unification (figure 1.2B). Thus, attributes trusted by the user (e.g., UniProt Accession Number) are used to merge and unify information across different databases. Although the use of this strategy is more time-consuming, it has the advantage that it is extensible to new data repositories, data types and attributes defined by the user.

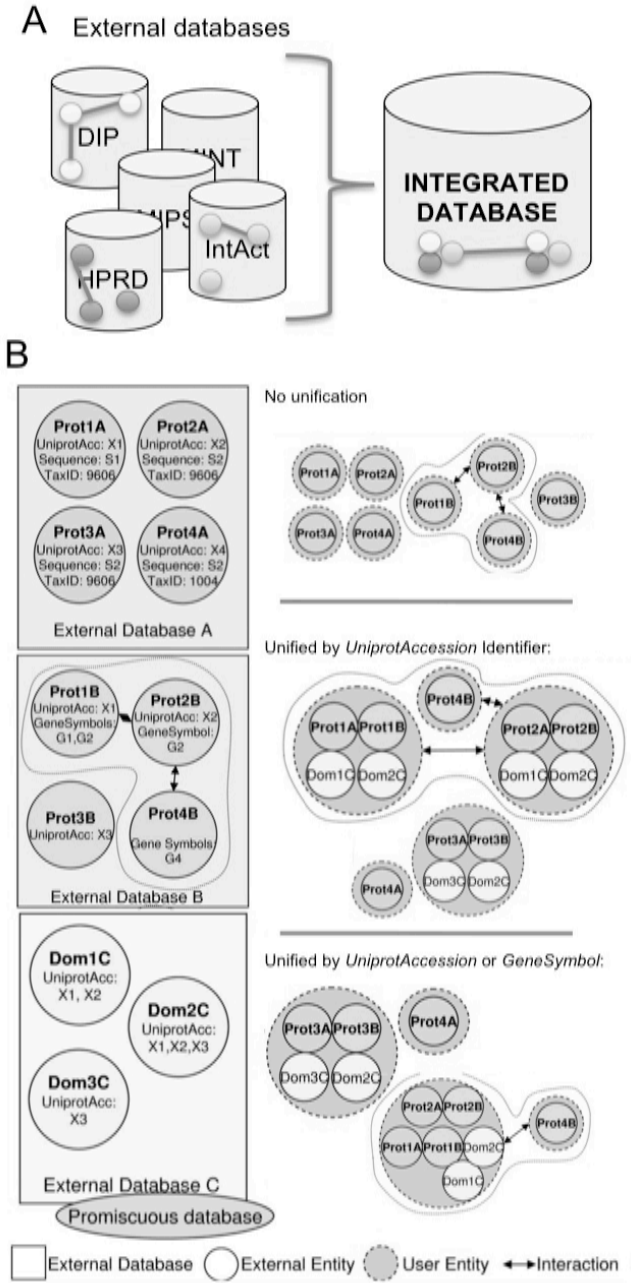


Figure 1.2. (Obtained from (98)) A. Data warehouse system. Several databases are parsed and stored in a single database. Equivalent entries are fused and redundancies are eliminated. Database access can be done through direct access to the database, by an application or by a web-server. B. BIANA user-driven integration examples. External databases provide different identifiers for their entities (external entities). According to the identifiers selected by the user, the integrated database can have different unified entities (user entities). Entities can be different biological molecules, such as proteins or genes, while relations can include any type of relationships, such as PPIs or common biochemical pathway among others.

1.2 Proteins and protein structure

Although the regulatory role of non-coding nucleic acids is currently being unravelled (99,100), proteins mediate most biological functions[5]. In fact, proteins are the bricks and mortar of cells. The work of proteins is structural and functional, as they are the principal element of the organization of the cell architecture, but they also play a relevant role in its metabolism and regulation. To perform all these functions, proteins need to adopt a particular three-dimensional (3D) structure (101-103). Thus, gaining knowledge about the 3D structure of proteins is crucial to understand their complex functions within the cell.

1.2.1 Protein structure

It is well-known that a protein's function is determined by its three dimensional (3D) structure (Thornton and cols. reviewed this issue in detail in (104)), which in turn is mainly dictated by its sequence (105). Actually, the amino acidic sequence of a protein constitutes its *primary structure*. Amino acids in the polypeptidic chain adopt recognizable structural patterns in the space, namely alpha helices and beta sheets (106). These regular patterns along with the more disordered regions that connect them conform the *secondary structure* of a protein. Higher levels of structural organization include *structural motifs* and *domains*. Structural motifs are defined as "simple combinations of a few secondary structure elements with a specific geometric arrangement that have been found to occur frequently in protein structures" (107). Domains represent an even higher level of structural organization, and are characterized by their capability of autonomously acquiring a 3D conformation (108). Domains have also been described as basic units of protein function (109). The global 3D structure of a polypeptidic chain (i.e. a protein) is known as its *tertiary structure*. Regarding protein complexes, the *quaternary structure* describes how components relate to each other, including atomic details of the interacting interfaces.

[5] The word 'protein' derives from the ancient Greek word 'protos', meaning first

First 3D structures of proteins were solved 50 years ago (110). Giving insights into proteins function, 3D structures have influenced the work of scientists in many areas of life sciences. However, the structural characterization of proteins is costly and time consuming, rendering the solution of the solution of all proteins 3D structure an unfeasible task. In order to surmount this problem, rapidly after a critical number of protein structures were made available Chothia and Lesk quantitatively measured the impact of amino-acid changes in closely related protein sequence on their structure. They concluded that similar sequences -this is evolutionary related proteins, exhibit nearly identical structures, and even distantly related proteins share the same fold (111,112). This is the very basis for comparative modeling, which aims to build a 3D model for a protein of unknown structure (the *target*) on the basis of the sequence similarity to proteins of known structure, usually referred as *templates*.

A final aspect to be considered about the structure of a protein is its plasticity. Structural elements in proteins are not static, but rather in permanent motion. This mobility is crucial to understand the protein function, especially if the protein binds small molecular ligands (113) (e.g. membrane receptors or enzymes) or other macromolecules such as proteins (114). Even more, some proteins are characterized for being intrinsically disordered (intrinsically unstructured proteins, IUPs). Often, such proteins participate in molecular interactions and their unstructured regions only acquire their functional conformation upon the presence of their ligand (115,116).

1.2.2 Protein structure determination

There exist several experimental techniques to elucidate the 3D structure of proteins. The following sections succinctly describe such techniques.

1.2.2.1 X-Ray crystallography

X-ray crystallography is based on the fact that when X-rays collide with electrons in matter, the beam of light is spread into many specific directions. The spread produces a specific pattern of scattered X-rays that is experimentally observed as a characteristic

electron density map when the atoms are ordered as in a crystal. This pattern is characteristic the atoms of the matter in which the X-rays collided. Hence, if X-rays were directed to a biomolecule such as a protein, its atomic structure could be determined. The scattering of an individual protein is very weak. In order intensify the signal, several instances of the protein must be arranged in a regular geometric pattern, exhibiting long-range order and symmetry producing a diffraction pattern. Such spatial organization is geometrically obtained in a lattice (an array of points repeating periodically in three dimensions) and physically in a crystal. To obtain a protein crystal, high concentrations of the protein (or the biomolecular complex) along with the appropriate experimental conditions are required (117,118). However, techniques for the production of small crystals and yet useful for protein structure determination purposes have been recently described (119).

Due to the mobile nature of proteins, not all crystals yield diffraction patterns suitable for the interpretation of the atomic details of the protein, even if crystallisation conditions are optimal. In such cases, the electron density maps obtained from the X-ray diffraction may not allow positioning the amino acid side chains although may suffice to trace the backbone of the protein or identify its fold (120). Furthermore, flexible regions such as loops may adopt different conformations in each cell of the lattice, causing an irregular dispersion of the colliding X-rays. As a result, the electron density map area corresponding to the flexible region of the protein appears blurred, phenomenon that may cause discontinuities in the final structural solution. Similar reasons hinder solving the structure of IUPs by X-ray. Additionally, it has to be noted that the functional conformation of a protein is not always the same conformation that the protein adopts in the crystal. Such conformational discrepancy may difficult the interpretation of the structural results. A similar problem appears with proteins that have several alternative conformations, which is a normal event in proteins that bind other molecules. In this case, all conformations could be present in the crystal, blurring the signal produced by the scattered X-rays. Despite all these drawbacks, X-ray crystallography has the unique advantage of being unrestricted by the protein size. Because of this, X-ray crystallography has been successfully used to determine the 3D structure of large macromolecular machines such as the proteasome (121) or the ribosome (122).

1.2.2.2 Nuclear Magnetic Resonance spectroscopy

Difficulties associated to X-ray crystallography, including not only its timely cost but primarily its problems to solve flexible regions or IUPs, has prompted the necessity of alternative methods for experimentally determining the 3D structure of proteins. Nuclear Magnetic Resonance (NMR) spectroscopy has become the most important of such alternatives. To cope with the aforementioned problems, NMR is applied to molecules in solution, which represents a more natural environment for most globular proteins. The method takes advantage of the fact that some atomic nuclei (including ^{1}H, ^{13}C, and ^{15}N) possess a magnetic moment (nuclear spin), which gives rise to different energy levels and resonance frequencies in a magnetic field. In MNR the magnetic field is induced from short pulses of electromagnetic (radiofrequency) energy, which prompts the raise of the energetic level. Excited nuclei return to their equilibrium state emitting radiation and resonating in a frequency that is characteristic of the atomic environment of the excited nuclei, the so called chemical shift. Pulses of different radio frequencies may provide different data about the environment. Two of the ore frequently used types are COSE (correlation spectroscopy), which by detecting covalent links allows to determine adjacent residues, and NOSEY (nuclear Overhauser enhancement spectroscopy), which provides information of residues closer in the 3D space, regardless of their positioning in the protein sequence (120).

NMR experiments yield spectra populated with numerous peaks, making the interpretation of the experiment results a hard task. Hence, it is not always possible to univocally assign a protein amino acid to a certain peak in the spectrum. Kurt Wüthrich and colleagues solved this problem in the early 1980 decade (123), and since then NMR has been successfully applied to solve the structure of proteins. However, the high complexity of the obtained spectra normally makes the technique only useful to relatively small sized proteins. Nevertheless, some complexes such as the GroEL-CroES chaperone have been analyzed using NMR based techniques (124).

1.2.2.3 Other methods for the structure determination of proteins

There also exist several other techniques for the determination of the 3D structure of proteins. These include cryo-electron tomography (125,126), small angle X-ray scattering (SAXS) (127), and solid-state NMR (128). Cryo-electron tomography is suited to the study of huge protein structures, such as the capsoids of virus (129) or the large macromolecular complexes that conform the cell machinery (i.e. the nuclear pore complex (130) or the cytoskeleton (131)). SAXS is applied to proteins in solution, being an appropriate technique to study IUPs (132). However, neither cryo-electron tomography nor SAXS can provide similar atomic details as X-ray crystallography or NMR do. Finally, solid-state NMR has become a useful tool to study the structure of membrane proteins at high-resolution rates (133).

1.2.3 Protein structure prediction

Despite the exponential increase of available sequences and 3D structures, the number of sequences highly exceeds that of 3D structures. This difference in numbers is proportional to the disparity of the costs for experimentally obtaining either the sequence or the structure of a protein. Therefore, covering the gap between sequence and structure becomes a compelling requirement to achieve a molecular understanding of the protein function. Theoretical methods can help to bridge this gap by inferring the 3D structure from the sequence. These methods are classified into three different groups: comparative modelling, fold recognition and new fold or ab initio methods.

1.2.3.1 Homology modelling

Homology or comparative modelling techniques are those devoted to infer the 3D conformation of a protein of unknown structure (*target*) from homologue proteins of known structure (*templates*). These methods are based on the assumption that structural features in proteins are more conserved that its sequences. Thus, two proteins with enough sequence similarity will fold in a similar way and share the same conformation in space (111,112). The process through which a tertiary structure is assigned to a given

sequence is carried out in three steps, namely: template identification, template alignment, and model building. Finally, the produced model should be assessed (134). Figure 1.3 summarizes the process of homology modelling.

Known 3D data of proteins is stored in the Protein Data Bank (PDB) (135). Thus, the identification of the template refers to the process of identifying the structure of the PDB whose sequence is the closest homolog of the target. Such sequence homology search can be performed using sequence alignment tools like BLAST and PSI-BLAST (136), or Hidden Markov Model (HMM) profile methods like HMMER (137), the two latter methods aiming to identify remote homologs of the target (138).

Once the template (or templates) has been selected, its sequence has to be aligned with that of the target. Depending on specific requirements, the alignments can be redone with other sequence alignment methods such as CLUSTALW (139) or T-COFFE (140). Additionally, some methods optimize the sequence alignment through a genetic algorithm protocol that iterates the alignment, model building and model evaluation in order to obtain the best possible alignment (141).

Model building is the process by which the three-dimensional data of the template(s) is applied on the query sequence. MODELLER is one of the most used and comprehensive modelling software (143). The program fits the target sequence onto the template structure upon satisfying a set of spatial constraints: (1) homology-derived constraints, (2) stereochemical constraints such as bond angles, and (3) statistical preferences for dihedral angles and non-bonded interatomic distances.

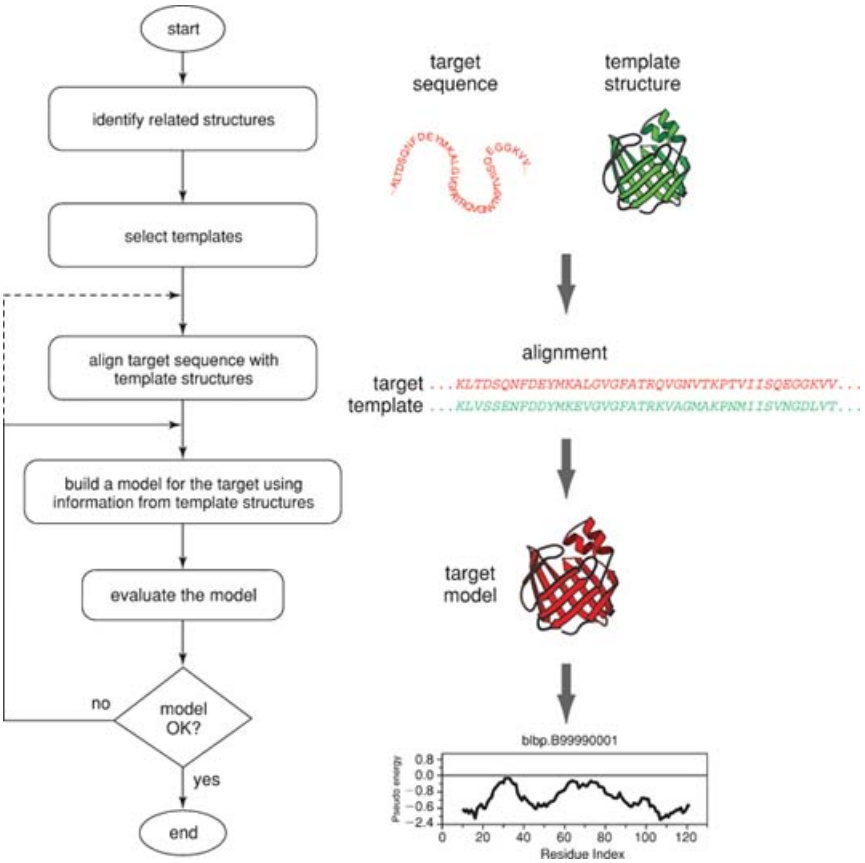


Figure 1.3. (Obtained from (142))Flowchart for single protein modelling.

1.2.3.2 Fold recognition

While comparative modelling relies on sequence similarity to infer the 3D structure of proteins, structural features can be used to detect remote sequence similarity. The approach is based on two different principles a) the observation that protein fold is better conserved than its sequence (111) and b) the fact that the number of structural folds that proteins adopt is limited (144,145). These facts have promoted the emergence of structural classifications to capture evolutionary relationships (146,147). The increasing number of initiatives in structural genomics (148,149) greatly boosts the odds that there exists a solved 3D structure of a protein with the same fold that a given target protein.

Early methods (150,151) recognized to which fold a protein belonged by "*finding sequences that are most compatible with the environments of the residues in the 3D structures*" (150), a technique that known as "threading". The probabilistic and

energetic description of such environments will be covered in this section. Newer methods introduce further variables to describe the environment, but are still restricted to predict an already known fold for a target sequence.

1.2.3.3 Ab initio *methods*

Beyond comparative modelling and fold recognition (which have been jointly termed "template based modelling") remains the challenge to predict an unknown fold for a protein sequence. One approach is to reproduce the twists, contorts and stretches that drive an amino-acidic chain from a theoretical thread-like shape to the fold the protein really has: its *native state* (152). This tortuous trail is known as the *folding path*. Successfully reproducing this path requires lots of computational efforts. Nevertheless, the combination of several theoretical frameworks allows an approximation to the problem using molecular dynamics (153). This combination is basically composed of:

i) the atomic or molecular degrees of freedom considered in the model,

ii) the definition of the forces that govern the system (*force field*) as a function of the chosen degrees of freedom,

iii) how these degrees of freedom are to be sampled and

iv) the boundaries of the system and the external forces that apply.

Applying molecular dynamics to the fold recognition problem is costly (and normally only applied to small peptides) due to the fact that the number of states in which a protein is unfolded is much greater than the number of folded states (154). Furthermore, using such *ab initio* approach holds the additional difficulty to recognize the native state, which does not necessarily correspond to the absolute energy minimum if relevant biomoloecular forces are not completely understood (155).

Baker and cols. faced this problem with a completely different approach (156). Analogously to how sequences can be threaded into folds, fragments of sequences can be threaded into fragments of different known structures, which can blend to new folds. As Cozzetto and Tramontano stated (157), *"this method was inspired by the observed*

local sequence-structure correlation in possibly unrelated proteins at different hierarchical levels, from a few residues to supersecondary structure elements" (158,159). This approach has been successfully used to pose highly accurate 3D models for single domain proteins (160). Interestingly it is currently being combined into a single strategy with template-based modelling (161). Only recently, advances in multi-domain proteins (a problem similar to the quaternary structure of proteins) have been reported (162).

1.2.4 Flexible regions

While most of regular structures in a protein maintain a semi-rigid pose that allows the identification of recognizable folds (146), flexible regions of proteins acquire ever-changing conformations, which make solving their structure a problematical task. Among the flexible regions of the proteins, *loops* –defined as regions between two regular structures, are of particular interest for describing a proteins function. Particularly, loops have been found to participate in forming binding sites and enzyme-active sites (107). It is noteworthy, that conformational differences between homologous proteins with different functions are known to occur often in the regions comprising turns and loops (163,164). Loops may occasionally be placed in the protein surface, and the model based on the Optimal Desolvation Area (ODA) (165) suggests that several regions of the protein, or even a high percentage of its surface, may be relevant for the molecular association of two proteins. It is in agreement with recent findings of Wass *et al*. who showed the use of docking experiments (see section 1.3.3.1) to identify the interacting partners of a protein as a consequence of implied restrictions in the protein surface conformation (166). Further more, it has been shown that loops are key elements to enable or disable the formation of PPIs (167).

Despite the fact that loops where classically regarded as random structure regions in the protein, (107), certain structural and geometrical patterns can be observed among them, allowing for a classification of loops (168,169). Such classifications can be exploited to identify function-associated loops in enzymes (170) or to identify interaction signatures in PPIs (see section 4).

1.3 Protein-protein interactions

The important role of proteins within the cell cannot be totally understood without grasping how they interact with other proteins and biomolecules (171). The molecular association of proteins is central for numerous cellular functions, including the formation of cell molecular machinery and the propagation of environmental signals. Hence, comprehending the complex network of protein interactions is a necessary means for understanding how do the cells work.

1.3.1 Protein interaction detection

Several methods have been developed identify physical interactions between two proteins. Protein Complementation Assays (PCA) (172) represent the group of most commonly used methods. In PCA protocols, the proteins of interest (*bait* and *prey*) are covalently linked at the genetic level to incomplete fragments of a third protein known as the *reporter*. Commonly, the reporter protein is a transcription factor that regulates a certain gene, which upon activation prompts an observable phenotype. The whole system is expressed in vivo. If bait and prey proteins interact, the reporter fragments are close enough to become functional, and consequently the reporter activity is detectable. Among PCA methods, the most widely used is the Yeast Two-Hybrid assay (Y2H), which has been widely used in low and high-throughput experiments (173).

Other PCA methods offer additional interesting features. For instance, MAPPIT (174) is a membrane-based PCA that upon reporter protein complementation allows the reconstruction of a membrane STAT (Signal Transducer and Activator of Transcription) used for the study of modification-dependent PPIs in mammalian cells. TOXCAT (175) is an alternative membrane-based PCA method that measures the association strength between protein transmembrane helices in biological membranes. Other PCA methods are based on cytoplasmic reporters, such as the Protein Kinase A fused with complementary fragments of a bioluminescent reporter (i.e. Renilla luciferase) (176).

Such methods are used to study the dynamics protein complexes (association and dissociation).

An other group of methods is aimed to detect weak and transient interactions an their location in living ells, and is based in fluorescence. Among these methods, the Green Fluorescence Protein complementation assay (GFP) (177) and the Bimolecular Fluorescence Complementation (BiFC) (178) are the most popular. Interestingly, these methods can be applied in high-throughput fashion while complemented with other techniques. Particularly, BiFC, which measures the interaction strength based on fluorescence intensity, can be combined with flow cytometry, providing a fast and highly sensitive method to validate weak protein interactions (179). Förster/fluorescence Resonance Energy Transfer (FRET) (180) is another widely used flourescence-based assay in which energy from a donor fluorophore can be transferred to an acceptor fluorophore if this is close enough and appropriately oriented. Bioluminescence Resonance Energy Transfer (BRET) (181) is even more sensitive than FRET; in this assay the the donor fluorophore is replaced by a luciferase.

Traditional proximity-based methods like PCAs involve the creation of fusion proteins between the targets (prey and bait) and a partial reporter. As an undesired side effect, the fusion of the target and the partial construct of the reporter may affect the binding ability of the targets. Aimed to surmount this problem, in situ Proximity Ligation Assays (PLA) detects PPIs without generating fusion proteins with high selectivity and sensitivity. In PLA two modified antibodies are ligated against the two target proteins; when the targets are in close proximity, the antibodies can emit a signal (182). Furthermore, PLA is able to provide the subcellular localization of PPIs in situ at single-molecule resolution.

A different family of methods for PPI detection is based on the array technology. These methods have several proteins are covalently attached to a planar support (*probes*), and their ability to interact specifically with other labelled proteins (*samples*) is measured (183). An interesting variation of array-based methods is the Surface Plasmon Resonance array (SPR) (184), in which the samples are non-labelled. Instead, an optical biosensor identifies the molecular binding events by detecting changes in the local refractive index, providing real-time affinity and kinetic data.

Finally, all the methods described in section 1.2.2 can be also used to detect PPIs, providing further information on the structural details of the interaction. Table 1.3 summarizes the methods commonly used to detect PPIs.

Table 1.3 (Adapted from (98)) Experimental methods commonly used to gather information related with protein protein interactions. The first column displays the name of the method, and the second the type of method. Third to fifth columns show if the method can detect binary interactions (Binary), multiple protein complexes (Complex) or if can be used in a high-throughput fashion (HT).

Method	Type	Binary	Complex	H.T.
Yeast Two Hybrid (Y2H)	PCA.	✓		✓
Mammalian PPI trap (MAPPIT)	PCA.	✓		
Tox-r dimerization assay (TOXCAT)	PCA.	✓		
Bimolecular Fluorescence Complementation (BiFC)	PCA. Fluorescence.	✓		
Proximity Ligation Assay (PLA)	PCA.	✓		
Förster/fluorescence resonance energy transfer (FRET)	Fluorescence.	✓		
Bioluminescence Resonance Energy Transfer (BRET)	Fluorescence.	✓	✓	
Protein microarrays	Array.	✓	✓	✓
Surface Plasmon Resonance Array (SPR)	Array.	✓	✓	
Tandem Affinity Purification (TAP)	PCA.	✓	✓	✓
X-ray crystallography	Other.	✓	✓	
Nuclear Magnetic Resonance (NMR) spectroscopy	Other.	✓	✓	
Cryo-electron tomography	Other.	✓	✓	

1.3.2 Protein interaction prediction

High-throughput methods have produced large amounts of PPI data, but their reliability and coverage has been questioned (185,186). Several computational methods have been developed in order to complement experimental techniques, and provide different levels of information detail.

1.3.2.1 Methods for predicting binary interactions

The prediction of binary interactions is the task of recognizing interacting partners regardless of the regions implied in the molecular association or the atomic details of the interacting interface. Computational approaches can be used to infer new predictions but also to validate, corroborate, or explain the experiments. Such methods can be grouped according to the basic hypothesis used for the prediction:

i) Genome based methods, such as domain fusion (187), gene neighborhood (188), or phylogenetic profiles (189);

ii) Experimental knowledge-based approaches, such as interologs (190), domain profiles (191) or sequence signatures (69);

iii) Methods based in evolution, such as correlated mutations (192) or phylogenetic mirror trees (193);

iv) Methods based on chemical properties of the dynamics of the interaction, such as prediction of kinetic rates for molecular association (194);

v) Docking techniques are normally employed to gain insight in the interacting region (see section 1.3.3.1), but have been used to infer protein partners in a pilot experiment(166).

1.3.2.2 Methods for predicting the interaction region or interface

The knowledge of the conformation of a binary complex formed by two proteins, at least in their interface, is essential to understand the molecular mechanisms involved in their docking (195). The first step in determining how a PPI is produced is to discover the regions involved in it. Computational methods aimed to solve this problem can be further divided into two categories, depending on whether they require knowledge of the interacting partners of the protein to be analysed or not.

The identification of interacting regions regardless of the protein partners is possible due to the fact that interface regions share specific features that distinguish them from the rest of the protein:

i) Residues in interface regions are highly conserved due to evolutionary constraints (196).

ii) PPI interfaces have shown to bear specific physico-chemical properties due to different amino-acid composition propensities (197-199).

iii) Interacting regions present structural constraints that can be measured in terms of Optimal Desolvation Area (ODA) (165).

iv) By combining different sources of information machine-learning methods can predict binding site (200,201).

In addition, the information about the interacting partners of a protein can provide information to identify its interacting region. For instance, the binding residues of a PPI are subject to co-evolution constraints (202,203). Also, structural and sequence patterns extracted from complexes with known structure have been used to predict interaction interfaces (204). Not all the residues participating in the interaction are equally important; the ones contributing more significantly to the binding free energy have been defined as hot-spots (204). Hot-spots are characterized by stronger structural (205) and evolutionary (206) constraints when compared to the rest of the protein. Finally, network topology based methods have also been successfully used to identify the binding regions in PPIs (73) (see section 1.1.5).

1.3.3 Details of the protein interaction interface

Beyond the identification of the interacting surface, a more detailed level of knowledge can be obtained studying the structural details of the residues forming the interaction. Precise molecular details of the protein complexes are available in the Protein Data Bank (PDB) (135) or in derived databases, such as PROTCOM (207), 3did (71), iPfam (208) or PRISM (209). However, there is a large gap between the number of protein

complexes with known structure and the total amount of known interactions and complexes. Predictive computational methods are key to reduce this gap (195), and fall into two different categories depending on the strategy they follow to predict such molecular details of the interacting region:

i) docking strategies, and

ii) comparative modelling strategies.

1.3.3.1 Docking approaches

Protein docking approaches are aimed to elucidate the structures of binary bio-molecules (e.g. two proteins) when experimental data regarding the structure of the complex is lacking but the structures of the interacting proteins are known. Docking methods sample the orientation of two unbound protein structures to produce several predictions about their interaction, followed by a scoring step to rank the predictions. These methods were introduced in 1978 (210). Since then, docking algorithms have substantially improved, with a breakthrough in algorithm speed given by the introduction of the Fast Fourier Transform (FFT) (211) (e.g. FTDock (212), ZDock (213), PIPER (214)), and by some other very successful geometry-based methods (e.g. FRODOCK (215), Hex (216), MolFit (211)). A docking procedure usually involves several steps (217). First, a rigid-docking search is performed by treating the two proteins as rigid bodies. One of the proteins, called the receptor, is kept fixed while the other protein, the ligand, is rotated and translated around the first. Next, further refinement of some structures takes place, allowing changes in conformation of the two unbound structures upon binding (114,218); this step may or may not be supported by experimental evidence. Docking algorithms return a large list of poses (bound conformations) that include many false interactions. Thus, the different docking poses need to be ranked by means of a scoring function. Two different types of scoring functions are used:

i) energy functions based on physical and chemical characteristics of the binding interface, such as ZRANK (219); or

ii) based on knowledge-based properties of known PPIs stored in structural databases, also known as statistical potentials (220,221).

1.3.3.2 Comparative modelling strategies

Comparative modelling strategies are based on the principles described in section 1.2.3.1. However, several automated pipelines to use comparative modeling to model the structure of macromolecular complexes in high-throughput have been developed. For example, the modelling automation provided by MODPIPE (222) and the resources of structural information provided in PIBASE (223), allowed Davis et al. (224) to apply homology modelling at a proteome scale. Also, Tuncbag et al. (209) has recently developed a protocol (based on PRISM (209)) for rigid-body structural comparisons, using the known structure of protein-protein interfaces and further flexible refinements. Fold-recognition based tools (see section 1.2.3.2) have been also developed for the prediction of the structure of protein complexes. Specifically, MULTIPROSPECTOR (225) and M-TASSER (226) are different methods that implement this multimeric threading. As a representative of *ab initio* techniques for protein structure prediction (see section 1.2.3.3), the Rosetta approach has been extended to solve the protein-protein docking problem (227). Once the structure of a PPI complex has been modelled, methods that use an atomistic description of the PPI (i.e. statistical potentials) can be used to assess its reliability. InterPreTS (228), which evaluates the reliability of a PPI based on the known structure of a homologous interaction, is one of such approaches. Finally, a recently developed analytical framework, SAPIN (229), combines the use of comparative modelling strategies, interaction specific statistical potentials, and empirical force fields (see section 1.2.3.3) to predict binary interactions after gaining detailed insights on the interacting region.

1.3.4 Unveiling the structure of large protein complexes

Available methods for the determination of the structure of proteins (see section 1.2.2) or for gaining insight into the interacting region (see section1.3.3) may be useful to

elucidate the structure of small complexes. However, the assembly of large macromolecular complexes such as the nuclear pore complex (7), which contains more than 450 proteins, requires a different approach termed *integrative modelling*. Such approach is based on the integration of the different sources of structural information available for different parts of the complex and its components. The main idea of this methodology is to use particular characteristics of the complex that can be synergistically combined in order to restrict the possible solutions to only those consistent with the available structural information.

Several low-resolution techniques for the structural determination of proteins, including cryo-electron tomography, SAXS (see section 1.2.2.3), or cryo-electron microscopy, can provide valuable information about the molecular shape of the complex. High-resolution structural information of some of the complex components and knowledge about the binary PPIs within the complex impose further restrictions in the orientation and positioning of the complex subunits. Then, fitting all the information together is similar to assembling a puzzle (see Figure 1.4). Multifit (230) and DOMINO (Discrete Optimization of Multiple Interacting Objects) (231) are computational methods designed to automate this task. This approach has been successfully used to obtain the structural model of the nuclear pore complex (7), the chromatin (8), or the RNA polymerase (232).

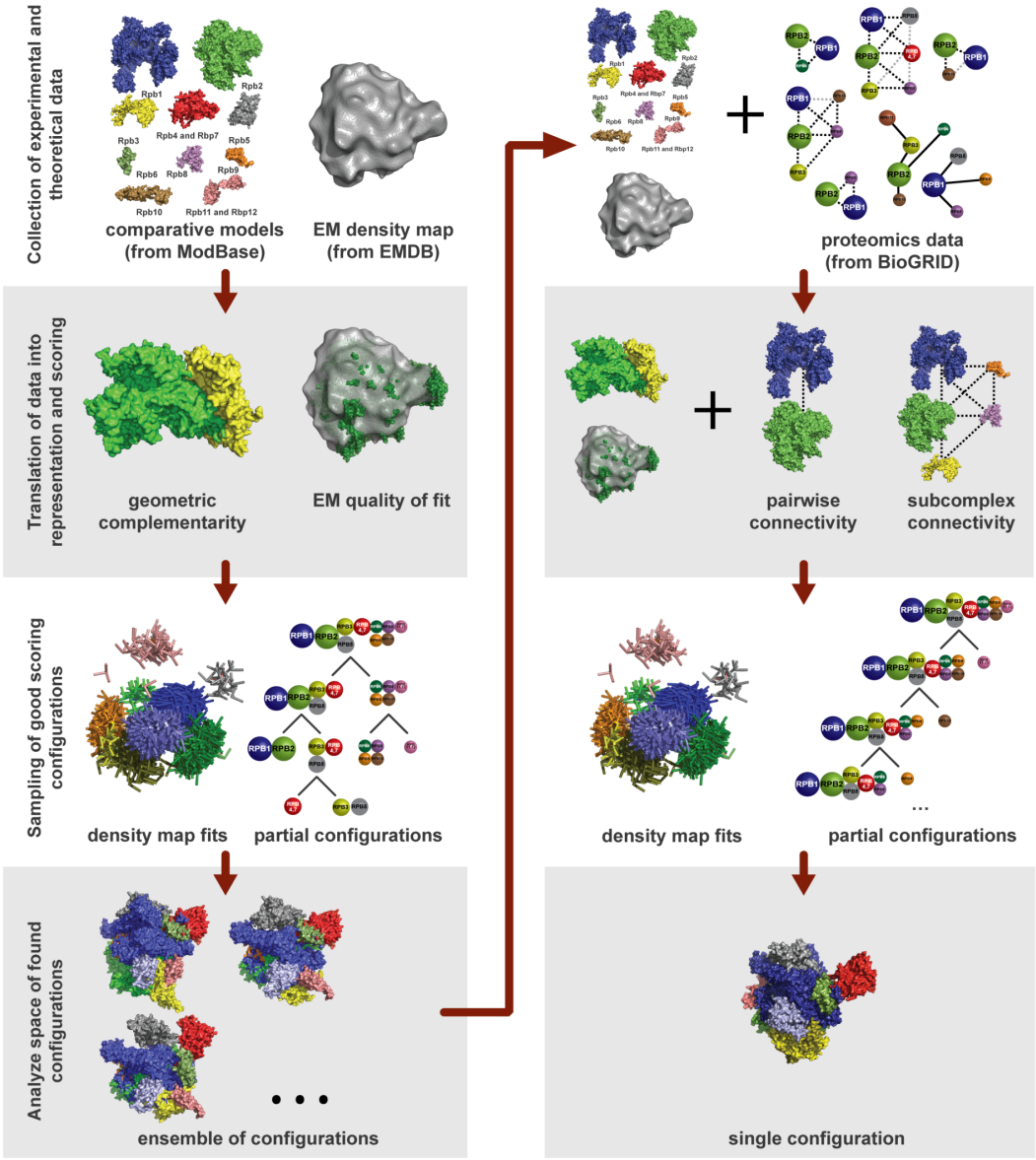


Figure 1.4. (Obtained from (233)) Schematic representation of integrative modelling of the human RNA polymerase II (232).

1.3.5 Protein interaction repositories

Results from PPI detection experiments and predictions are deposited in public repositories of biological interactions, which enable a convenient access to the information available and facilitate its further analysis. Table 1.4 summarizes the principal resources providing information on PPIs. However, as previously stated (see section 1.1.6) the integration of such data in unified systems is still a challenging task.

As a result, a single PPI could be represented several times in repositories, hindering the performance of computational methods that rely on known PPIs to make their predictions. Furthermore, incompleteness (false negatives) and noisiness (false positives) of available PPI data complicates the interpretation of such data, for instance in terms of yielding estimates of the interactome size (234,235). Negative data (i.e. protein pairs known not to interact) could be used to surmount these problems.

Table 1.4 (Adapted from (98)). Data repositories for PPI data. Databases are grouped according tot the nature of data they enclose. The level of detail each database provides is encoded in the third column according to the following legend: 1 binary PPIs; 2 interface region; 3 structural detail.

Databases	Information	Level(s) of detail
STRING (68)	Functional relations between proteins (not necessarily PPIs) inferred using genome-based methods and literature text-mining (see section 1.3.2.1)	1
BIND (236), IntAct(237), DIP (238), BioGRID (239), HPRD (240), MINT (241), MPact (242), MIPS (243), HPID (244)	Complex composition and protein binary pairs determined experimentally.	1
PIPs (245), OPHID (246), POINT (247)	Predicted PPIs obtained with different methods	1
Domine (248), PSIbase (249)	Domain-domain interaction pairs observed in PDB database	1, 2
PCRPi-DB (250), HotRegion (251), HotSprint (252), ASEdb (253)	Residues found in the interface region accounting for the majority of the binding energy, also known as hot spots.	2
iPfam (208), 3DID (71), SCOPPI (254), SCOWLP (255), PIBASE (223), InterPare (256), PRINT (257).	Structurally determined domain-domain interaction (DDI) interfaces.	2, 3
PDBSUM (258), PROTCOM (207)	Databases of protein complexes.	3
Protein-protein docking benchmark (259)	High-resolution structures that are non-redundant at the family level and for which the structure of each unbound interacting partner is also known.	3
InterEvol (260)	Evolution of protein complex interfaces.	2, 3

1.3.6 Negative protein-protein interaction data in the context of PPI prediction

The use of negative sets (i.e. pairs of proteins which are known to not interact) in the development of interaction discovery or prediction techniques is clear: when testing the efficacy of any approach, having "gold standard" sets of both positives and negatives is critical (261). A common approach for defining negative PPI datasets, exploits the fact that proteins from different cellular locations are unlikely to interact (262). However, this approach leads to a bias on the estimation of the accuracy of predictive methods, since additional constraints related to localization render the prediction task easier (263). Another option widely used is to employ random datasets (261,264-266). This approach may lead the predictor to learning the pattern of missing values and, thus, cause an over-prediction of associations (267). Furthermore, current estimates indicate that for each 1000 protein pairs, only 1 of them actually interacts (65,234,268). While this rate implies a low risk for enclosing real interactions in randomly generated negative models, such risk may be unacceptable for certain tasks (e.g. the study of interaction specificity between two protein families), or can be increased by the imposition of functional restrictions in the random models (i.e. proteins with similar functional annotation) (269).

On this context, the use of experimentally tested negative data would be the most convenient choice to construct a negative "gold standard". Despite the fact that PCA methods for PPI detection (see section 1.3.1) may yield relevant data about non-interacting pairs (270), very few data about actual non-interacting pairs (NIPs) has been compiled. The Negatome database (271) constitutes a recent effort to catalogue such information and contains about two thousand negative interactions, half of them derived from manual literature curation and the other half from the analysis of 3D structures of protein complexes. The main criticism to the Negatome database is centered in its scale limitation and its evident experimental bias. Only recently, a method to exploit the negative PPI information obtained from PCA methods for PPI detection has been developed (269).

1.4 Motivation of this thesis

Proteins constitute the brick and mortar of living cells, being the main responsible for most biological activities of the cell. To fulfil this duty, most proteins associate with others, forming complexes that range from binary interactions to an impressive cell machinery. In other words, proteins rarely act alone; they rather constitute a mingled network of physical interactions and other types of relationships. In this context, understanding the function of a protein implies to recognize the members of its neighbourhood and to grasp how they associate, even at the atomic level. Unravelling these associations with experiments is expensive and time consuming. Hence, *in silico* predictions and network biology represent a convenient alternative to study protein-protein interactions.

Acknowledging the importance of the relationship between the sequence, the structure, and the function of a protein, the study and prediction of the structure of proteins arises as a crucial topic for understanding their molecular associations. Furthermore, the role of small local structures such as loops in the formation of protein-protein interactions has been widely hint, but not exploited yet for predictive purposes. In this scenario, the combination of structural information with the wide knowledge available about interacting proteins –the interactome, may hold unique prospects for the identification of new protein interactions.

Besides the interactome itself, other biological networks crucial for the cell survival such as signalling networks, rely on interacting proteins to fulfil their functions. Characteristic properties of protein interaction networks such as centrality and modularity can be used to predict new members of signalling pathways. Several methods have been developed to exploit such properties with predictive purposes; however the incompleteness and noisiness of PPI data makes this task difficult. Assessing the success of such methods in a controlled experiment may shed new light on the task of transferring the function of known participants in signalling networks to other candidate proteins yet unknown.

2 Objectives

The objectives of this thesis are:

i) Exploit the functional relevance of loops in the protein binding process to develop a new method for protein-protein interaction prediction.

ii) Study the potentiality of protein interaction networks as a tool to transfer functional annotation in a selected example of signalling networks: the apoptosis pathways. Particularly, this objective is focussed in exploiting the results from different methods used for functional annotation transferring in a controlled experiment to achieve more reliable predictions.

The first objective has been accomplished by developing *iLoops*, a new method to predict protein-protein interactions. The method is based in the observation of characteristic loop signatures in known interacting and non-interacting protein pairs. In order to make the iLoops method available to the scientific community a web server tool has been done. The two manuscripts included in section 4 treat this objective and have been recently submitted to scientific journals. The second objective is considered in section 5 and has been tackled by using different methods to transfer annotation from 53 well-studied members of the human apoptosis pathways (as known by 2005) to their protein interactors. The results obtained by the different methods were compared to the knowledge gained on the apoptosis pathways in the period 2005-2010. Taking advantage of this retrospective approach, a scoring function was developed to select the most reliable candidates for the apoptosis pathways. The results from this study were published during year 2012. Additionally, two review publications about the prediction of tertiary and quaternary structure of proteins were produced as a result of the research that lead to the accomplishment of the stated objectives. These publications are included in section 3.

3 Results

3.1 Prediction of 3D structure of proteins and protein complexes

Centeno, N. B., Planas-Iglesias, J. & Oliva, B. (2005). Comparative modelling of protein structure and its impact on microbial cell factories. *Microb Cell Fact* **4**, 20.

Planas-Iglesias, J., Bonet, J., Marín-López, M. A., Feliu, E., Gursoy, A. & Oliva, B. (2012). Structural Bioinformatics of Proteins: Predicting Tertiary and Quaternary Strucutre of Proteins from Sequence. In *Protein-protein interactions. Computational and experimental tools.* (Cai, W. & Hong, H., eds.). InTech, Rijeka.

3.1.1 Comparative modelling of protein structure and its impact on microbial cell factories

Centeno, N. B., Planas-Iglesias, J. & Oliva, B. (2005). Comparative modelling of protein structure and its impact on microbial cell factories. *Microb Cell Fact* **4**, 20.

3.1.2 Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence

Planas-Iglesias, J., Bonet, J., Marín-López, M. A., Feliu, E., Gursoy, A. & Oliva, B. (2012). Structural Bioinformatics of Proteins: Predicting Tertiary and Quaternary Strucutre of Proteins from Sequence. In *Protein-protein interactions. Computational and experimental tools.* (Cai, W. & Hong, H., eds.). InTech, Rijeka.

3.2 Prediction of protein interactions based on local structural features

iLoops manuscript:

Planas-Iglesias, J., Bonet, J., Garcia-Garcia, J., Marín-López, M.A., Feliu, E. & Oliva, B. (2012) Understanding protein-protein interactions using local structural features Under second revision in *Journal of molecular biology*.

iLoops web server manuscript:

Planas-Iglesias, J., Bonet, J., Marín-López, M. A. & Oliva, B. (2012). iLoops: A protein-protein interaction prediction server based on local structural features *Submitted to Bioinformatics*.

3.2.1 Understanding protein-protein interactions using local structural features

iLoops manuscript:

Planas-Iglesias, J., Bonet, J., Garcia-Garcia, J., Marín-López, M.A., Feliu, E. & Oliva, B. (2012) Understanding protein-protein interactions using local structural features Under second revision in *Journal of molecular biology*.

Understanding protein-protein interactions using local structural features

Joan Planas-Iglesias[1], Jaume Bonet[1], Javier García-García[1], Manuel A. Marín-López[1] , Elisenda Feliu[2] and Baldo Oliva[1*]

[1] Structural Bioinformatics Group (GRIB-IMIM). Departament de Ciències Experimentals i de la Salut. Universitat Pompeu Fabra. C/ Dr. Aiguader 88, 08003 Barcelona, Catalonia, Spain

[2] Department of Mathematical Sciences. University of Copenhagen. Universitetsparken 5, 2100 Copenhagen, Denmark

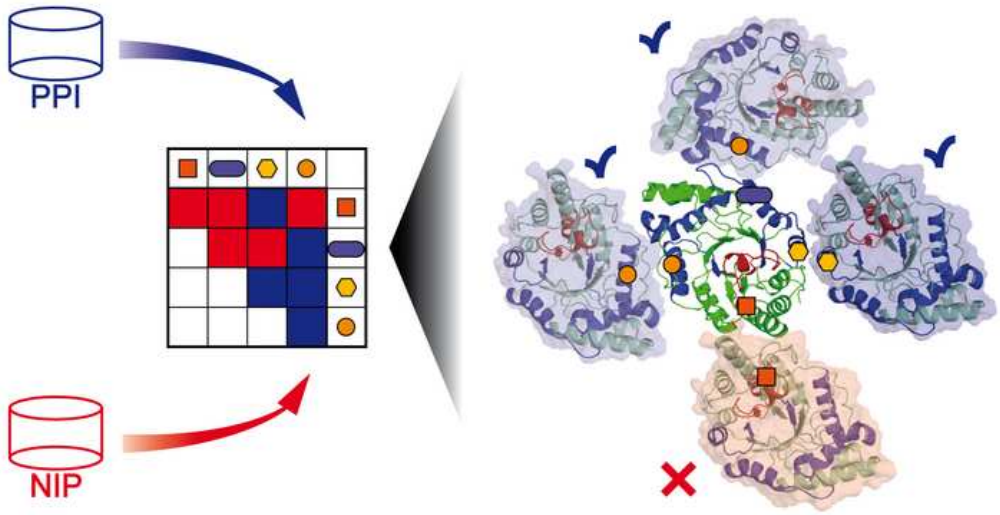
Keywords: Protein-protein interactions**;** funnel-like theory; protein interaction prediction; functional loops; Negatome database.

Abbreviatures: PPI protein-protein interaction; NIP non-interacting pair; BIANA Biological Interactions and Network Analysis; PRS Positive Reference Set; NRS Negative Reference Set; PES Positive Evaluation Set; NES Negative Evaluation Set.

[*] **Corresponding Author.**

Baldo Oliva, baldo.oliva@upf.edu

Graphical Abstract



Abstract

Protein-protein interactions play a relevant role among the different functions of a cell. Identifying the protein-protein interaction network of a given organism (interactome) is useful to shed light on the key molecular mechanisms within a biological system. In this work, we show the role of structural features (loops and domains) to comprehend the molecular mechanisms of protein-protein interactions. A paradox in protein-protein binding is to explain how the unbound proteins recognize each other among a large population within a cell and how they find their best docking interface in a short time-scale. We use interacting and non-interacting protein pairs to classify the structural features that sustain the binding (or non-binding) behaviour. Our study indicates that not only the interacting region but also the rest of the protein surface is important for the interaction fate. The interpretation of this classification suggests that the balance between favouring and disfavouring structural features determines if a pair of proteins interacts or not. Our results are in agreement with previous works and support the funnel-like intermolecular energy landscape theory that explains protein-protein interactions. We have used these features to score the likelihood of the interaction between two proteins and to develop a method for the prediction of PPIs. We have tested our method on several sets with unbalanced ratios of interactions and non-interactions to simulate real conditions, obtaining accuracies higher than 25% in the most unfavourable circumstances.

Introduction

Protein-protein interactions (PPIs) are crucial to understand how proteins perform their cellular functions[1]. However, due to the limitations in the experimental methods for determining PPIs and the structure of protein complexes, there is a large gap between our knowledge on genome sequences and the discovery of PPIs. To diminish this gap, several techniques such as two-hybrid assays and

affinity purifications followed by mass spectrometry have afforded large-scale identification of PPIs, producing a vast amount of data during the last decade[2; 3]. Simultaneously, several repositories have stored PPIs (reviewed inTuncbag *et al.*[4]) and other tools have been developed to integrate this information in order to exploit all available relationships[5]. While the majority of these efforts provide long lists of interacting proteins, they still miss the molecular information on the interface regions involved in the interactions. Therefore, completing protein interactome maps and understanding how proteins interact are still milestone challenges in current biology[6].

Computational methods for PPI detection represent a feasible alternative to experimental methods. Genomic-based approaches such as sequence homology and phylogenetic profiling, co-evolution, and co-localisation may identify interacting pairs, but further structural knowledge is required to unveil the interface between two proteins[7]. Homology modelling techniques may address this issue If structural templates are available[8]. On the lack of templates, docking approaches (recently reviewed in Janin *et al.*[9]) are designed to predict the conformation of protein-protein interactions. Nevertheless, docking yields thousands of solutions that are challenging to rank[10]. Valencia and co-workers recently showed that sets of docking poses can be used to discern between interacting and non-interacting protein pairs, although the native conformation may be indistinguishable among the numerous docking solutions[11].

Sprinzak and Margalit pioneered the work that correlated pairs of domains with protein-protein interactions[12]. Being basic units of protein folding and function, protein-domains have been widely used to exploit PPIs since then, either to identify new interactions or to define the interaction region[13]. Nevertheless, remote homologues with identical structural domains (i.e. conserved fold structure) may have different protein functions. This functional diversity is often associated with variations in the protein surface that occasionally can take place in the loops (regions connecting two elements of secondary structure)[14]. Hence,

it is not surprising that protein loops are associated with protein functions and can be used to predict protein annotation or to enable and disable interactions between single-domain proteins[15].

It has been proposed that the variety of well-ranked docking poses may reveal the possibility of near-native solutions, whose further optimization can recover some of the loosened contacts[16]. This hypothesis supports the concept of the funnel-like intermolecular energy landscape used to describe PPIs[17; 18]. The principle behind this hypothesis is that when two proteins collide they recognize their potentiality to interact, even if the interface produced in this approach is not optimal.

We have based our study in the characterisation of deterministic structural features of interacting and non-interacting pairs[19]. Using the Negatome database[20] and integrating PPIs from several databases we have classified pairs of structural features (defined as interaction signatures) that are characteristic of PPIs or Non Interacting Pairs (NIPs). Our results strongly suggest that it is the balance between interacting and non-interacting structural features that determines if a pair of proteins will interact or not. Taking advantage of this result, we have developed a method to predict protein-protein interactions that may be really valuable for experimentalists.

Results

Assignment of classified loops and domains to proteins.

We obtained two reference sets of pairs of proteins -one for PPIs and the other for NIPs- with some structure associated (pairs of proteins to which we could assign standardised loops or domains, see Methods). These sets were referred as the Positive Reference Set (PRS) and the Negative Reference Set (NRS), respectively. Local structural features, namely SCOP domains[21] or loops classified in ArchDB[22], were assigned to proteins in these sets (see

supplementary Table S1). We characterised the proteins from these sets with *protein signatures,* groups of up to three structural features: domains, loops, or loops belonging to the same domain, denoted as {L}, {D}, and {L_D} respectively (see Methods).

Pairs of protein signatures

Protein signatures were used to define the *interaction signatures* of protein pairs. An interaction signature was defined as the combination of two protein signatures of the same type ({L}, {D}, or {L_D}), one from each partner. Interaction signatures were denoted as *positive* if observed in the PRS and as *negative* if observed in the NRS (see Methods). We calculated the number of times an interaction signature appeared in protein-pairs of the PRS or the NRS (Figure 1). The total number of interaction signatures and protein signatures are summarized in supplementary Table S1.

Overview of the classification of PPIs using local structures (interaction signatures)

We expected that over-represented interaction signatures were characteristic of their reference set. Hence, we assigned to each interaction signature a p-value (equation (1) in Methods) corresponding to the probability of observing the interaction signature at least as often as the number of occurrences in the reference set. This probability follows a hypergeometric distribution (with values between 0 and 1). For the sake of comparison with the work of Sprinzak and Margalit[12], we also used their log-odds based score and applied it to domain interaction signatures (see equation (2) in Methods).

These scores aim to statistically evaluate the occurrence of one interaction signature. However, several interaction signatures (derived from the PRS, the NRS, or both) can be assigned to protein pairs. Taking this into account, we designed six different classifiers for each type of signature to classify the

interaction between two proteins (see details in Supplementary Note 1). These classifiers are: the lowest p-value among the positive and negative interaction signatures found in a protein pair (pV^+ and pV^- respectively), the total number of positive and negative interaction signatures in the pair (S^+ and S^- respectively), the logarithm of the ratio between the negative and positive lowest p-values (*LpVR*), and the logarithm of the ratio between the number of positive and negative signatures (*LSR*). In addition, we denoted by LO^+ and LO^- the log-odds scores (equation (2) in Methods) calculated with interaction signatures and protein signatures in the PRS and NRS, respectively. Similarly, we denoted by *LOR* the difference between LO^+ and LO^-.

To analyse the ability of the above classifiers to discern between PPIs and NIPs, we used a five-fold cross-validation approach (see Methods and Supplementary Note 2). The score of an interaction signature was computed from its frequency in the training set. For each classifier, we built Receiver Operating Characteristic (ROC) curves for all test sets and averaged the Area Under the Curve (AUC) as a measure of their suitability and performance. We restricted the training and test sets to protein pairs with less than 40% sequence identity in order to avoid over-training (see Supplementary Note 3).

Finally, we designed separating functions combining the number of interaction signatures and the p-values to classify pairs of PPIs and NIPs. The separating functions combined S^+ and pV^+, S^- and pV^-, or *LSR* and *LpVR* (see Supplementary Note 4).

Performance of the classifiers derived from positive and negative interaction signatures

The classifiers derived from the PRS and NRS were used to discern between PPIs and NIPs. The basic principle underlying the prediction of PPIs using pV^+ or the prediction of NIPs using pV^- was the presence of a highly specific signature

for one or the other. Nevertheless, the fact that a protein pair enclosed a large number of positive interaction signatures –regardless of their p-value, could also suggest a potential interaction. This was the rationale behind the application of the S^+ classifier. The same rationale using negative signatures was applied for the S^- classifier to predict NIPs. We have to note that the results presented in this and the following sections correspond to the analysis performed using non-homologous reference sets (see Supplementary Note 3). The analyses of the results using all sequences are provided in Supplementary Notes 5 and 6.

In Table 1 are shown the AUCs of these classifiers and in the supplementary Table S3 the associated errors. It was noteworthy that pV^- and S^- classifiers achieved better AUCs than pV^+ and S^+. This result shows the potential of negative interaction signatures to identify NIPs. Besides, using the 5-fold approach with non-homologous sequences of the PRS worsened the ability of pV^+ and S^+ classifiers to identify PPIs (see Supplementary Note 5, Table 1 and supplementary Table S2). This was not observed with the pV^- and S^- classifiers. We argue that homologs of a pair of interacting proteins have some probabilities to preserve the interaction, while homologs of non-interacting pairs have the same chances to interact (or not) as any other pair.

We also observed that PPIs had a large number of positive signatures (S^+) and one or more signatures with low pV^+. Therefore, we designed hyperbolic separating functions of pV^+ and S^+ to discern between PPIs and NIPs (see Supplementary Note 4). With these functions we obtained a PPV_{PPI} around 55% and PPV_{NIP} around 75%. Similar trend was observed for NIPs using S^- and pV^- classifiers, obtaining 65% PPV_{PPI} and above 95% PPV_{NIP} (see supplementary Table S4 and Methods section for definitions).

Analysis of log-ratio classifiers: LpVR and LSR

The principle for the use of the log-ratio approach was to identify which type of signature (positive or negative) was more relevant in terms of p-value (LpVR) or

in the number of signatures (LSR), in order to decide if a pair of proteins could interact. For the sake of comparison, we also studied the log-odds classifier *LOR* (the difference between LO^+ and LO^-). We compared the results between all classifiers to guarantee the selection of the best criteria to distinguish PPIs and NIPs (see Table 1, Figure 2, and supplementary Table S3).

We examined if separating functions of *LpVR* and *LSR* for each type of signatures could discern between pairs of PPIs and NIPs. We intuited from the plot of *LpVR* versus *LSR* that linear separating functions could suffice to separate PPIs and NIPs (see Figure 3 and Supplementary Note 7). The PPV_{PPI}, and the NPV_{PPI} (i.e. PPV_{NIP}) were around 0.8 (see Table2 and refer to Methods for definitions of PPV_{PPI}, and NPV_{PPI}).

Assessment of the prediction of PPIs.

In the previous sections we proved the ability of some derived classifiers to distinguish between PPIs and NIPs on a one to one ratio. However, an experimentalist faces a different ratio when testing the interactions of a random selection of protein pairs of a proteome. Usually the number of NIPs is much larger than the number of PPIs and only the expertise of the user can help to reduce this difference. Therefore, to assess the predictive power of these classifiers in real conditions, we needed a new set of PPIs and NIPs where the number of NIPs was higher than the number of PPIs. We extended the sets of PPIs and NIPs and called the new sets Positive Evaluation Set (PES) and Negative Evaluation Set (NES), respectively. Details on the construction of these sets are shown in Methods.

To evaluate each type of interaction signature ({L}, $\{L_D\}$, and {D}) we split the PES and NES in three subsets: one to obtain interaction signatures, another to train a random forest classifier using the WEKA package[23], and the last one to test the prediction. In order to avoid trivial predictions we removed protein pairs with more than 40% sequence identity before splitting into groups (see

Supplementary Notes 3 and 8 for details). We used a 1:1 ratio between PPIs and NIPs to obtain positive and negative interaction signatures and train a random forest classifier. However, we used unbalanced ratios between PPIs and NIPs to test the prediction (1:10, 1:20 and 1:50) in order to simulate the problem faced by an experimental biologist when predicting PPIs. In the last step of our approach, we penalized the errors due to false positive predictions from the random forest classifiers with different "relative costs" (Methods). The values of TPR were almost unaffected by these relative costs. Figure 4 and Table 3 show the average of PPV obtained with different "relative costs" and unbalanced ratios of PPIs over NIPs using {L} interaction signatures (results for $\{L_D\}$ and {D} interaction signatures are shown in supplementary Tables S5 and S6, respectively).

We have to note that the set of co-localized proteins represents around 5% of all possible protein pairs in a human cell. Thus, the 1:50 ratio simulates the naturally occurring unbalance between PPIs and NIPs in the human proteome, which has been estimated to be about 1 PPI for 1000 NIPs[24]. Consequently, the ratios of our test represent different levels of expertise of a user that applies the prediction method: from an ideal expert, who is able to select pairs with a 50% probability to have an interaction, to a non-expert, who almost randomly selects any pair of proteins of the interactome with ordinary good judgement (see further in Supplementary Note 8).

Comparison with previous works

First, we assessed if the p-value (equation (1) from the main text) was equivalent to the log-odds score in equation (2). Similarly to the work of Sprinzak and Margalit[12], we neglected sequence similarity and predicted PPIs using pV^+ and LO^+ using positive interaction signatures of domains {D}. We compared the ROC curves (see Figures 2C and 2D), their AUCs (supplementary Table S2) and the associated estimation of the error of AUC (see supplementary Table S3). The results using pV^+ were comparable to those obtained with the log-odds score used by Sprinzak and Margalit[12] to predict PPIs.

Also, we compared our predictive approach with other methods extracted from the literature. We compared the baseline hypothesis of each method and the construction of the benchmarks employed for each validation test (most differences arose from the construction of the set of NIPs). The majority of methods achieved recalls of more than 80% but neglected to report the ratio of success (PPV) considering that the number of NIPs should be larger than the number of PPIs. We wish to note that any of the compared methods were tested with similar realistic conditions to the ones applied in this study. Table 4 summarises the principal features and results.

Examples

We illustrate positive interaction signatures using the interaction between tsunagi (RBM8A_DROME) and the mago nashi (MGN_DROME) proteins of *Drosophilla melanogaster*. The structure of this complex is given by the PDB code 1oo0[25] (Figures 5A and 5B). These two proteins are components of the core of the *Exon Junction Complex* (EJC) that plays a key role in the localisation of "oskar mRNA", thus being vital for the development of the fruit fly. The classifiers LpVR and LSR, where $pV^+ < pV^-$ and $S^+ > S^-$, show that this as an example of a correctly predicted interaction.

To illustrate negative interaction signatures we selected the complex of the *Rab geranylgeranyltransferase* (RabGGTase α and β subunits) and their accessory *Rab Escort Protein* (REP) (PDB code 1ltx[26]). REP is vital for the recognition of proteins in the Rab family, and the post-transcriptional modifications generated by RabGGTase are crucial for the reversible membrane association that Rab requires to function. REP is recognised by α-RabGGTase, having no physical contact with β-RabGGTase (Figures 5C and 5D). The prediction by means of LpVR and LSR is that both proteins should not interact ($pV^+ > pV^-$ and $S^+ < S^-$).

The prediction of the interactions in the human exosome is provided in Supplementary Note 9 and supplementary Figure S1 in order to show the limits of our approach.

Discussion

We have explored the capability of several structural features to explain the mechanism underlying the formation of protein-protein interactions. Besides structural domains, we used the classification of loops, defined as combinations of two contiguous secondary structures, with the same purpose. We developed a scoring method to score the likelihood of a protein-protein interaction using groups of loops and domains. These groups were defined as protein signatures for each protein. We characterised pairs of protein-signatures derived from PPIs as positive interaction signatures. Compared to previous scoring methods[12], our score provided the probability of observing the interaction signature at least as often as the number of occurrences in a reference set. We corroborated recent findings about the crucial role of loops in the formation of PPIs[13; 27], but also in preventing the formation of protein interaction complexes. We found that negative interaction signatures discerned between PPIs and NIPs, showing that relevant information was enclosed in NIPs.

Scoring the likelihood of an interaction based on the p-values of positive and negative interaction signatures considered only one interface produced by a putative collision. Thus, we analysed whether the total number of signatures could be a good criterion of classification. Our results may be interpreted in the light of a recent study made by Wass *et al.*[11]. In their work, they studied the capability of docking algorithms (which are used to identify the best interface between a pair of proteins) to predict interacting partners regardless of their success in pinpointing the interacting region. Their study showed that although docking algorithms could fail to identify the native complex (and thus, the

interface), the distribution of docking scores discerned between interacting and non-interacting pairs. From their results, the authors suggested that protein surface morphology contained sufficient information to identify a *bona fide* interactor[11]. This concept implied that several regions of the protein were important for the molecular association between two proteins. Indeed, this described a model of the funnel-like intermolecular energy landscape in PPIs[17; 18]. It is unclear how two interacting proteins can find each other within a large population of proteins and quickly form a binary complex. If the molecular association was the result of a quasi-infinite series of elastic collisions with a unique successful outcome (i.e. the final pose of the binary complex), the formation of PPIs would require an unaffordable time-scale (NP-problem) as in Levinthal's paradox[28]. The solution of the problem is to assume that in the collision of two proteins they recognise if they have to interact or not, forming an intermediate complex that may (or may not) have the best docking interface. For a non-interacting pair, both proteins would be immediately released to interact with others, while if they had to interact they would stay together (or near each other) until finding the correct conformation. This model implies certain "stickiness" between the interacting proteins, which would allow the formation of the intermediate complexes.

In this context, our results suggest a similar explanation for the formation of interacting protein-pairs. We proved that the number of interaction signatures is a good classifier of PPIs and NIPs, suggesting that not only one interacting region is important to decide whether a pair of proteins could interact. Hence, several protein regions could participate in the interaction process, allowing the formation of intermediates of the binary complex. In the framework of the funnel-like intermolecular energy landscape theory, proteins would explore their energetic landscape during the protein-protein collisions. We propose that this landscape is constrained by the composition of local structural features of both proteins (protein signatures). According to this model, positive and negative interaction signatures could represent energetic valleys and peaks respectively. Thus, the

pairing of protein signatures would encode the possibilities to accept or not the interaction, as illustrated in Figure 6. A large number of positive signatures would increase the probability to find a positive signature in the first collision, which in turn would help to maintain the partners close. With a similar argument, a large number of negative signatures would release both proteins unbound. Also, if the best positive-signature score were better than the best among the negative signatures, the probability to retain the interaction of the two proteins would be larger, and vice-versa. Finally, a protein-protein interaction is a state that results from the many occurring collisions between molecular pairs. Thus, two proteins interact or not depending on the balance between the log-ratio classifiers *LpVR* and *LSR,* which were used to predict their potential binding.

Interestingly, the explanation is valid for all types of protein signatures, either loops or domains. This suggests two different scales in 3D space, one for large proteins formed by several domains, and another for small or single domain proteins. In the case of large proteins formed by several domains, a large number of positive (or negative) signatures imply several domains and groups of them favouring (or hindering) the interaction. For small proteins formed by few domains or single-domains, local structures formed by pairs of secondary structures, and/or groups of them, play this role. It is also noteworthy that classifications using {L} and $\{L_D\}$ signatures obtain similar results. This implies that {L} and $\{L_D\}$ signatures cannot be distinguished, suggesting that most loop signatures playing a role in the decision to accept or deny an interaction belong to the same domain. In short, domains would play the main role to decide if a pair of proteins interacts in a low resolution scale, while at higher resolution scale the best pair of interacting domains would be decided by the selection of loop interaction signatures ({L}).

We proved the application of this approach to predict PPIs under different unbalanced ratios between the number of PPIs and NIPs, trying to simulate the conditions of real experiments. For a certain ratio, the system allows us to

address different putative questions, such as predicting the largest amount of real interactions or minimizing the number of failures. Therefore, our approach wraps up a unique framework for experimental biologists who want to predict PPIs and require prioritising sets of candidate pairs. Our approach provides probabilistic expectations for the set of proteins-pairs according to the knowledge of the user about them and his expertise to select the best candidates (around 30% of success in very unfavourable conditions or more than 80% of recall in the best scenario).

To summarise, we have shown that loops and domains appropriately describe the interactions between proteins and, once grouped into interaction signatures, they can be used to predict them. Furthermore, not only the likelihood of observing a particular interaction signature is important to determine whether two proteins will interact but also the total number of signatures plays a major role. Our findings clearly support the funnel-like intermolecular energy landscape theory for PPIs. Finally, we have constructed a method of prediction of PPIs under different conditions that may be worth for an experimentalist.

Methods

Experimental Datasets

We used BIANA[29] to integrate data of PPIs from several repositories (HPRD[30], MINT[31], BioGrid[32], IntAct[33], and MIPS[34]). Protein entries from different databases were unified if they shared a UniProt Accession Number, a GeneID or had the same protein sequence. Pull-down experiments were avoided and we studied only direct binary interactions. A total of 117970 PPIs conformed this set.

A dataset of NIPs was obtained from the Negatome database[20]. We used the manual-stringent dataset, which contains 1162 NIPs extracted from individual experiments reported in scientific literature (high-throughput experiments were not considered).

Assignment of loops and domains.

We assigned loops and domains to each protein in the PPI and NIP sets. Protein loops were defined as classified in ArchDB[22] and protein domains were defined as classified in SCOP[21]. We annotated 2821 PPIs with domains and 632 PPIs with loops from ArchDB in the PRS. Due to the limited size of the NRS, we used sequence similarity to annotate loops and domains. First, we searched homologs in PDB using BLAST[36]. The hits used for the annotation had to satisfy a minimum percentage of identity according to the length of the alignment (above the twilight-zone curve, as described by Rost[37]). Second, we required a minimum sequence coverage of the structure (100% for loops and 75% for domains). We annotated structural features for 720 non-interacting protein pairs (699 with domains, 309 with loops and 288 with loops and domains).

The sets of PPIs and NIPs contained pairs of proteins such that both partners had assigned structural features of the classification of ArchDB, SCOP, or both. These sets were named Positive Reference Set (PRS) for PPIs, and Negative Reference Set (NRS) for NIPs.

Protein and interaction signatures

We defined as *protein signature* any group of up to three local structural features. We considered three different types of local structural features. Groups of ArchDB loops were named *loop signatures*, which were denoted by {L}; groups of SCOP domains were named *domain signatures*, which were denoted by {D}; and groups of ArchDB loops located in the same SCOP domain were denoted by $\{L_D\}$. For a pair of proteins (A,B), we defined an *interaction signature* as a pair of protein signatures of the same type, one from protein A and the other from protein B (see Supplementary Note 1).

Scoring Functions

We defined the *p-value* of a protein signature pair as:

$$M=\frac{N(N+1)}{2}; \quad n=P_i \cdot P_j; \quad pValue=\sum_{k=P_{ij}}^{k=\min(n,R)} \frac{\binom{n}{k}\binom{M-n}{R-k}}{\binom{M}{R}} \qquad (1)$$

where *i* and *j* are protein signatures from two different proteins; (i,j) is the pair observed in both proteins (either in PPI or NIP); *N* and *R* are respectively the total number of proteins (or domains for $\{L_D\}$ signatures) and the total number of pairs in the reference set (either PRS or NRS); P_i and P_j, are the number of observed proteins (for {L} and {D} signatures) or domains (for $\{L_D\}$ signatures) with the signature *i* or *j* respectively; and P_{ij} is the number of pairs with the interaction signature *(i,j)* in the reference set. The p-value is derived from the probability of observing the interaction signature in the reference set, which follows a hypergeometric distribution.

We also used the score defined by Sprinzak and Margalit[12] to associate pairs of domains with PPIs. This is calculated with the *log-odds* ratio between observed and expected frequencies of interaction signatures:

$$\log_2\left(P_{ij}/P_iP_j\right) \qquad (2)$$

Where P_i, P_j and P_{ij} are defined as above.

Five-fold cross-validation

We used a five-fold cross-validation approach to evaluate the capacity to distinguish between PPIs and NIPs with the classifiers defined in the text (Results section and Supplementary Note 1). The definition of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) varied according to the goal of the prediction (predicting PPIs or NIPs). Positives are putative interacting pairs and negatives are putative non-interacting pairs

when predicting PPIs, and viceversa when predicting non-interactions (see Supplementary Note 2). To assess the prediction capacity of the classifiers we computed the True Positive Rate (TPR), the False Positive Rate (FPR), the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV):

$$\begin{aligned} TPR &= TP/(TP+FN) \\ FPR &= FP/(FP+TN) \\ PPV &= TP/(TP+FP) \\ NPV &= TN/(TN+FN) \end{aligned} \qquad (3)$$

We used the sub-index PPI and NIP to identify the set of positives (i.e. PPV_{PPI} is the positive predictive value for predicting interactions and PPV_{NIP} is the positive predictive value for predicting non-interactions, therefore $PPV_{PPI} = NPV_{NIP}$ and $PPV_{NIP} = NPV_{PPI}$).

We computed the ROC curves and the AUCs with the ROCR package[38] five times, using different test sets. The difference between the extreme values of the AUCs and the AUC of the mean ROC curve was given as an estimation of the error of these averages. Average ROC curves and AUCs were obtained by averaging the results of the five tests.

Construction of the Positive and Negative Evaluation Sets.

The requirements of a good negative model for PPI predictions have been recently discussed[39]. On one hand, it is still nowadays difficult to identify non-interacting protein pairs due to the lack of sensitivity of high-throughput experimental detection methods of PPIs[24; 40]. On the other hand, it was shown that random negative models introduced biases[19; 41]. We used the Negatome database[20] as a negative reference set to derive the structural features preventing the interaction between two proteins. The Negatome data set had the advantage of taking into account co-expression and co-localization of the proteins forming a non-interaction pair. For instance, if two proteins are never co-

expressed or co-localised, they certainly would not have a chance to interact *in vivo*. Evidently, these proteins would not require encoding information to avoid the interaction. Hence, it would not be expected to learn negative interaction signatures from such a pair of proteins. On the contrary, coexisting pairs of non-interacting proteins might need to show repulsive features to be pulled apart. Therefore, it was worth to include them in the negative reference set. This rationale is an extension of previous findings of Ben-Hur and Noble[41], who showed that excluding co-localised protein pairs from the negative reference set introduced a bias for PPI predictions. We further used this criterion to extend the set of NIPs for the evaluation of the classifiers on the prediction of PPIs.

The Positive Evaluation Set (PES) was obtained by annotating loops from ArchDB and domains from SCOP in the set of curated PPIs by means of sequence similarity as described above. We were able to annotate domains and loops to 8207 and 7264 PPIs, respectively. To extend the set of NIPs we needed to define some conditions ensuring that a pair of proteins would not interact. First, we considered all proteins of the previous sets PRS and NRS and generated all possible pairs. Next, we removed all PPIs and any pair that could be predicted to interact by means of similarity using BIPS[42] (Supplementary Note 3) with a non-restrictive criteria (40%ID sequence similarity). We also ensured that the proteins of the pair were co-localized (sharing the same cellular component GO terms[43]). We obtained with this protocol 21155 pairs of proteins with unreported interactions. Finally, loops and domains were annotated for all the protein-pairs by means of sequence similarity (as described above) and pairs without structural features were removed. We were able to annotate domains of SCOP for 20229 protein-pairs and loops of ArchDB for 3361 pairs. This set was named Negative Evaluation Set (NES).

Acknowledgements

This work was supported by the grant FEDER BIO2011-22568 from the Spanish Ministry of Science and Innovation (MICINN) and by EU grant EraSysbio+

(SHIPREC) Euroinvestigación (EUI2009-04018). EF is supported by the postdoctoral grant "Beatriu de Pinós" from the Generalitat de Catalunya and the project MTM2009-14163-C02-01 from the ”Ministerio de Ciencia e Innovación". JPI, JGG and MAML acknowledge support by “Departament d'Educació i Universitats de la Generalitat de Catalunya i del Fons Social Europeu” through FI fellowships. JB is supported by BIO08-0206 grant from MICINN. Authors are thankful to Carsten Wiuf for his expert statistical advice, and to Mireya Plass and Attila Gursoy for useful discussion and critical proofreading of the manuscript.

Author’s contributions

BO, JPI, and JB conceived and designed the analyses and wrote the manuscript. JPI programmed, made the analyses and prepared the line art figures. JB implemented the scoring systems and prepared the rest of figures. JGG provided the unified protein interaction sets and performed the independent evaluation of the method. EF designed the scoring system. MAML automated the analysis pipeline. All authors read and approved the final manuscript.

References

1. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7.
2. Devos, D. & Russell, R. B. (2007). A more complete, complexed and structured interactome. *Curr Opin Struct Biol* **17**, 370-7.
3. Vidal, M., Cusick, M. E. & Barabasi, A. L. (2011). Interactome networks and human disease. *Cell* **144**, 986-98.
4. Tuncbag, N., Kar, G., Keskin, O., Gursoy, A. & Nussinov, R. (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform* **10**, 217-32.
5. Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T. P. & Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nat Methods* **6**, 75-7.

6. Petrey, D., Fischer, M. & Honig, B. (2009). Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc Natl Acad Sci U S A* **106**, 17377-82.
7. Aloy, P. & Russell, R. B. (2006). Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* **7**, 188-97.
8. Kundrotas, P. J., Zhu, Z., Janin, J. & Vakser, I. A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proc Natl Acad Sci U S A* **109**, 9438-41.
9. Janin, J., Bahadur, R. P. & Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Q Rev Biophys* **41**, 133-80.
10. Feliu, E. & Oliva, B. (2010). How different from random are docking predictions when ranked by scoring functions? *Proteins* **78**, 3376-85.
11. Wass, M. N., Fuentes, G., Pons, C., Pazos, F. & Valencia, A. (2011). Towards the prediction of protein interaction partners using physical docking. *Molecular systems biology* **7**, 469.
12. Sprinzak, E. & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of molecular biology* **311**, 681-92.
13. Sprinzak, E., Altuvia, Y. & Margalit, H. (2006). Characterization and prediction of protein-protein interactions within and between complexes. *Proc Natl Acad Sci U S A* **103**, 14718-23.
14. Russell, R. B., Sasieni, P. D. & Sternberg, M. J. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *Journal of molecular biology* **282**, 903-18.
15. Akiva, E., Itzhaki, Z. & Margalit, H. (2008). Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc Natl Acad Sci U S A* **105**, 13292-7.
16. Blundell, T. L. & Fernandez-Recio, J. (2006). Cell biology: brief encounters bolster contacts. *Nature* **444**, 279-80.
17. McCammon, J. A. (1998). Theory of biomolecular recognition. *Curr Opin Struct Biol* **8**, 245-9.
18. Tsai, H. H., Reches, M., Tsai, C. J., Gunasekaran, K., Gazit, E. & Nussinov, R. (2005). Energy landscape of amyloidogenic peptide oligomerization by parallel-tempering molecular dynamics simulation: significant role of Asn ladder. *Proc Natl Acad Sci U S A* **102**, 8174-9.
19. Park, Y. & Marcotte, E. M. (2011). Revisiting the negative example sampling problem for predicting protein-protein interactions. *Bioinformatics* **27**, 3024-8.
20. Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D. & Ruepp, A. (2010). The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res* **38**, D540-4.
21. Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* **36**, D419-25.

22. Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F. X., Sternberg, M. J. & Oliva, B. (2004). ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic acids research* **32**, D185-8.
23. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **11**.
24. Venkatesan, K., Rual, J. F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K. I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A. S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabasi, A. L. & Vidal, M. (2009). An empirical framework for binary interactome mapping. *Nat Methods* **6**, 83-90.
25. Shi, H. & Xu, R. M. (2003). Crystal structure of the Drosophila Mago nashi-Y14 complex. *Genes Dev* **17**, 971-6.
26. Pylypenko, O., Rak, A., Reents, R., Niculae, A., Sidorovitch, V., Cioaca, M. D., Bessolitsyna, E., Thoma, N. H., Waldmann, H., Schlichting, I., Goody, R. S. & Alexandrov, K. (2003). Structure of Rab escort protein-1 in complex with Rab geranylgeranyltransferase. *Mol Cell* **11**, 483-94.
27. Danielson, M. L. & Lill, M. A. (2010). New computational method for prediction of interacting protein loop regions. *Proteins* **78**, 1748-59.
28. Levinthal, C. (1968). Are there pathways for protein folding? *J Chem Phys*, 44-45.
29. Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J. & Oliva, B. (2010). Biana: a software framework for compiling biological interactions and analyzing networks. *BMC bioinformatics* **11**, 56.
30. Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T. K., Chandrika, K. N., Deshpande, N., Suresh, S., Rashmi, B. P., Shanker, K., Padma, N., Niranjan, V., Harsha, H. C., Talreja, N., Vrushabendra, B. M., Ramya, M. A., Yatish, A. J., Joy, M., Shivashankar, H. N., Kavitha, M. P., Menezes, M., Choudhury, D. R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C. K., Prasad, C. K., Kumar-Sinha, C., Deshpande, K. S. & Pandey, A. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic acids research* **32**, D497-501.
31. Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L. & Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic acids research* **35**, D572-4.
32. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research* **34**, D535-9.
33. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R. & Hermjakob,

H. (2007). IntAct--open source resource for molecular interaction data. *Nucleic acids research* **35**, D561-5.
34. Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H. W., Ruepp, A. & Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832-4.
35. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology* **215**, 403-10.
37. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering* **12**, 85-94.
38. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-1.
39. Yu, C. Y., Chou, L. C. & Chang, D. T. (2010). Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics* **11**, 167.
40. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Roncari, L., de Smet, A. S., Venkatesan, K., Rual, J. F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P. & Vidal, M. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods* **6**, 91-7.
41. Ben-Hur, A. & Noble, W. S. (2006). Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* **7 Suppl 1**, S2.
42. Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J. & Oliva, B. (2012). BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference. *Nucleic Acids Research* **In press.**
43. GeneOntologyConsortium. (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res* **38**, D331-5.
44. Jang, W. H., Jung, S. H. & Han, D. S. (2012). A Computational Model for Predicting Protein Interactions based on Multi-Domain Collaboration. *IEEE/ACM Trans Comput Biol Bioinform*.
45. Bjorkholm, P. & Sonnhammer, E. L. (2009). Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics* **25**, 3020-5.
46. Lee, H., Deng, M., Sun, F. & Chen, T. (2006). An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics* **7**, 269.
47. Maetschke, S. R., Simonsen, M., Davis, M. J. & Ragan, M. A. (2012). Gene Ontology-driven inference of protein-protein interactions using inducers. *Bioinformatics* **28**, 69-75.
48. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. & Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A* **104**, 4337-41.

49. Finn, R. D., Marshall, M. & Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21**, 410-2.
50. Raghavachari, B., Tasneem, A., Przytycka, T. M. & Jothi, R. (2008). DOMINE: a database of protein domain interactions. *Nucleic Acids Res* **36**, D656-61.
51. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci* **23**, 358-61.
52. Bader, G. D., Betel, D. & Hogue, C. W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-50.
53. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J. & von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561-8.

Figure Legends

Figure 1. Protein and interaction signatures.
A) Pairs of proteins (A,B). Domains are shown in big shapes and loops in small. A protein pair (A,B) is indicated with a connection symbol. B) The number of domain interaction signatures {D} is presented in a table (as in Sprinzak and Margalit[12]) using protein-domain signatures. C) Table showing the number of loop interaction signatures {L} and $\{L_D\}$. Protein signatures are formed by groups of 1, 2 and 3 loops from any domain, {L}, or from the same domain, $\{L_D\}$. Each cell of the table contains the number of {L} interaction signatures (in white background) and, if available, the number of $\{L_D\}$ interaction signatures (in grey background). When the number of $\{L_D\}$ and {L} interaction signatures is different, the value of {L} interaction signatures is highlighted in red.

Figure 2. Performance of classifiers using a 5-fold cross-validation.
Average ROC curves (TPR versus FPR) using interaction signatures {L} (in A and E), $\{L_D\}$ (in B and F), and {D} (in C, G and D). Figures 2A, 2B and 2C show the ROC curves using classifiers based in the p-value: pV^+ (in blue), pV^- (in red) and *LpVR* (in green). Figures 2E, 2F and 2G show the ROC curves using the classifiers in the number of interaction signatures: S^+ (in blue), S^- (in red) and *LSR* (in green). Figure 2D shows the ROC curves using the classifiers LO^+ (dark blue), LO^- (dark red), and *LOR* (dark green). Average ROC curves were obtained with protein pairs with sequence identity smaller than 99% (continuous line) and protein pairs with less than 40% sequence identity (dashed line).

Figure 3. Classification of PPIs and NIPs using LpVR and LSR classifiers.
Plot of *LpVR* versus *LSR* of PPIs (blue) and NIPs (red) calculated with interaction signatures {L} (A), $\{L_D\}$ (B), and {D} (C). The optimal line separating PPIs from NIPs is depicted in green.

Figure 4. Prediction of PPIs using random forests classifiers.
Averaged PPV (blue lines) and TPR (yellow lines)***** are shown as functions of the relative cost applied in WEKA package[23] using interaction signatures {L} (A), {L_D} (B), and {D} (C). Standard deviations are shown in error bars. Unbalanced ratios of PPIs versus NIPs are shown in different hues of blue: 1:1 (dark), 1:10 (navy), 1:20 (light), and 1:50 (cyan).
*****Note: TPR values were similar for all different unbalance ratios tested, and are shown only for the most unfavourable (1:50) since it encompass the largest error.

Figure 5. Examples of interaction signatures in structures of protein pairs.
Regions of loops involved in positive interaction signatures are shown in blue, while regions in negative interaction signatures are in red. The structure of the interaction between RBM8A_DROME (green) and MGN_DROME (wheat) is shown in figures A (positive interaction signatures) and B (negative interaction signatures). The complex between the RabGGTase (α subunit in black, β subunit in green) and REP protein (wheat) is shown in figures C and D. Loop interaction signatures of $\beta-$ RabGGTase and REP are shown in C (positive) and D (negative).

Figure 6. Schematic representation of protein signatures in protein pairs.
Proteins (A, B, and C) are represented as polyhedra and grouped in pairs (A,B), (A,C), and (B,C). Depending on whether they interact or not, proteins in each pair are connected by a black straight line or a red cross respectively. Faces in the polyhedra represent protein signatures, coloured in blue or red depending on whether they favour the interaction or the non-interaction. The interaction signatures formed by pairs of protein signatures of each protein are considered favourable or unfavourable according to the ratio of the positive and negative scores. Protein A can interact with Protein B because both have favourable protein signatures for this interaction; on the contrary, the interaction A-C is highly unfavourable due to their protein signatures composition. However, the

same protein signatures in C that made the interaction A-C unlikely, favour the interaction between B and C.

Table Legends

Table 1. AUCs of PPI classifiers.

Columns 2 to 10 show the average AUCs obtained with a five-fold approach for different classifiers trained and tested with non-homologous protein pairs (less than 40% sequence identity). The first column shows the type of signature used for the classification (Sign.).

Table 2. Classification of PPIs and NIPs.

Ratios of correctly classified PPIs (TPR_{PPI}), NIPs (TPR_{NIP}), and positive predictive values for the classifcation of interactions (PPV_{PPI}) and non-interactions (PPV_{NIP}) using linear separating functions of *LpVR* and *LSR*. The first column indicates the type of interaction signature of the classification.

Table3. Performance of the prediction of PPIs.

Averaged PPV and TPR of the prediction of PPIs using {L} interaction signatures and random forests classifiers (standard deviations are shown between parenthesis). Columns 2-6 indicate the results for different unbalanced ratios of PPIs versus NIPs and the first column indicates the relative-cost of false-positives versus false-negatives applied in the random-forest classifier.

*****Note: TPR values were similar for all different unbalance ratios tested, and are shown only for the most unfavourable (1:50) since it encompasses the largest standard deviation.

Table 4. Comparison of PPI prediction methods.

We compare several approaches by method, validation sets and statistical results. The type of method and input are shown in columns 2 and 3. The database (or approach) employed to construct the training and test sets are shown in columns 4 to 7 (4 and 5 for PPIs and 6 and 7 for NIPs). In column 8 (%ID) we indicate the maximum percentage of sequence identity between the sequences of the training and tests sets (99% means that this restriction is not

applicable and test and training sets contain very similar sequences). The balance ratio between PPIs and NIPs is shown in column 9 (Ratio). Statistical values for the comparison are shown in columns 10 to 16 (sensitivity, specificity, PPV, accuracy, F1 measure, and AUC). The symbol "-" indicates that the values could not be collected from the original work. We compared our approach based in loop interaction signatures with 3 methods based on domain-domain interactions (rows 1-3), one method based on Gene Ontology (row 4) and methods based on the protein sequences (rows 5-7). References for these methods are indicated in column 1 (use them for further details on the methodology and benchmarking databases). For the description of methods: "Correlation" indicates methods based in the correlation between specific features (such as domains or loops) with PPIs; "MLE+Bayesian" indicates the Maximum Likelihood Estimation and Bayessian networks; "SVM" is for Support Vector Machines; "Shortest path" is the distance in the gene ontology tree; and "Net. Weight" is a network-weighted approach. For the description of databases we use the same names as in the main text or from the original works: "9DIN" stands for the integration of nine Domain Interaction Networks (see reference in the table); "Y2H&HPRD" is the integration of yeast-two-hybrid experiments and human interactions from HPRD (see reference in the table). Additional databases iPfam[49], DOMINE[50], PQS[51], BIND[52], HPRD[30], and STRING[53] were used to train and test the methods. The majority of methods used a random selection of pairs of proteins for the set of non-interactions (identified as "Random"), while we specifically selected the set of non-interactions for the evaluation (NES).

*Note: results are given for an effective cost 1:20.

Figure 1

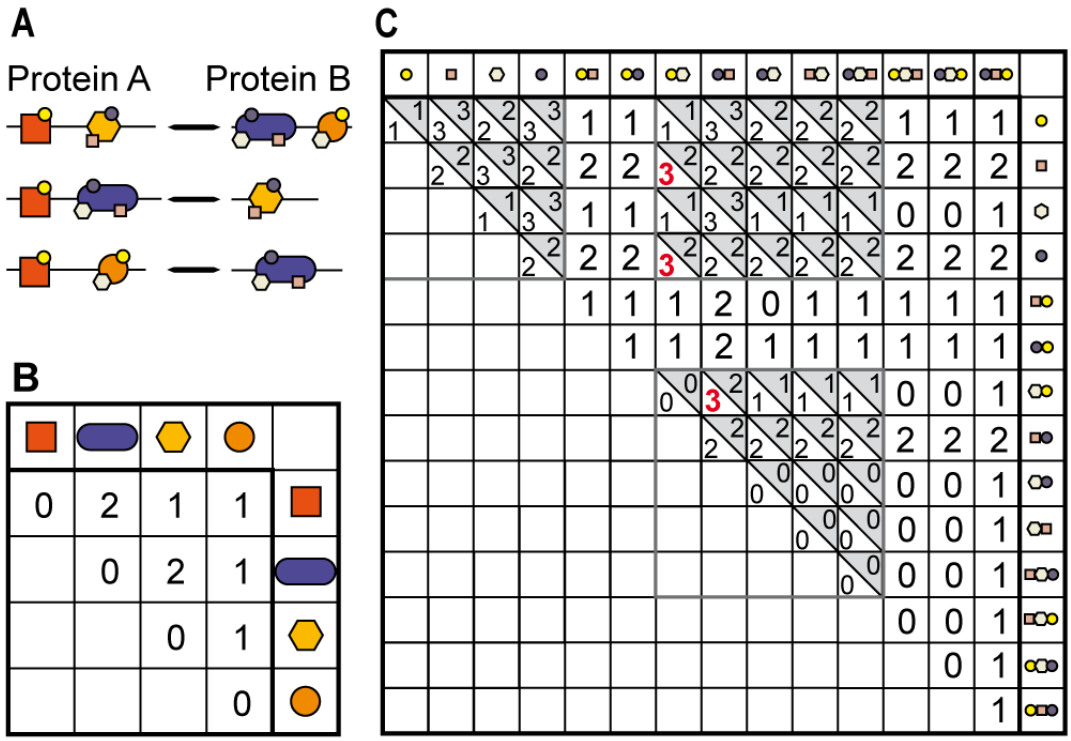


Figure 2

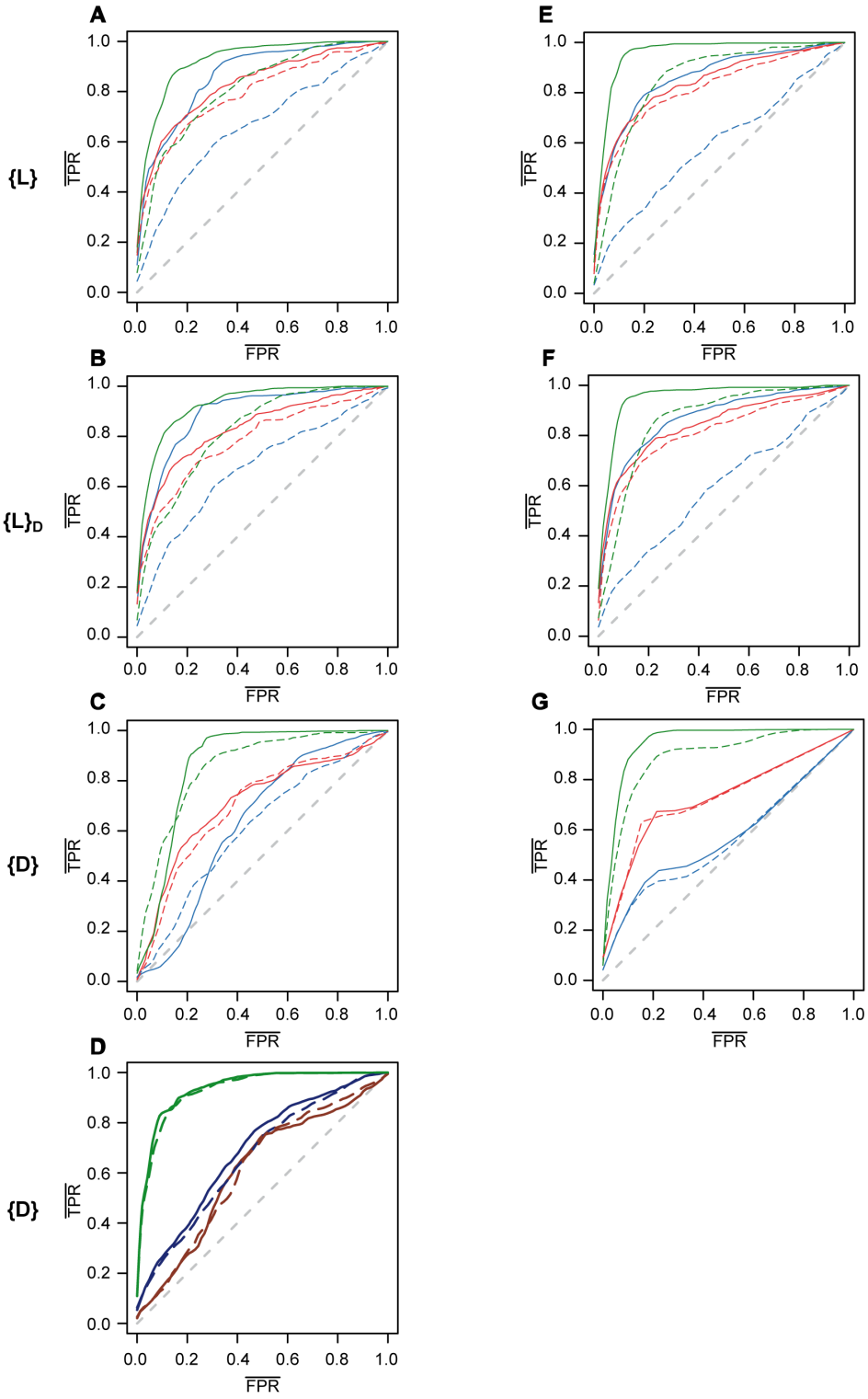


Figure 3

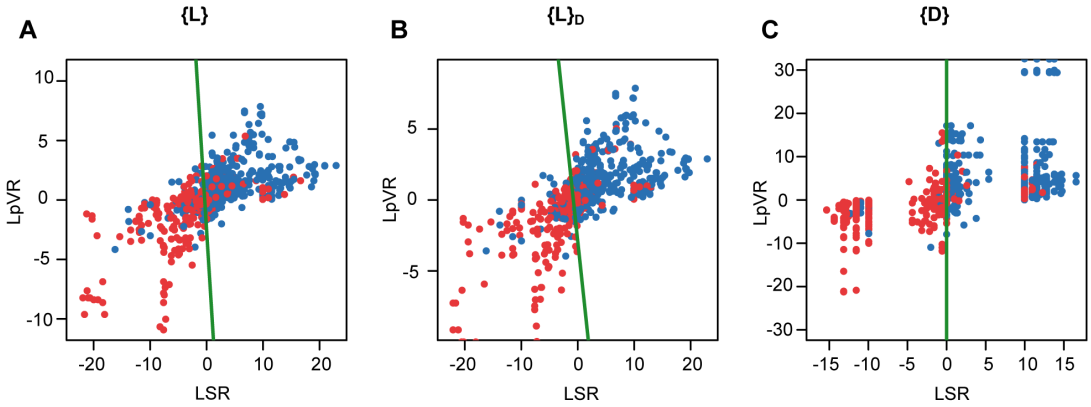


Figure 4

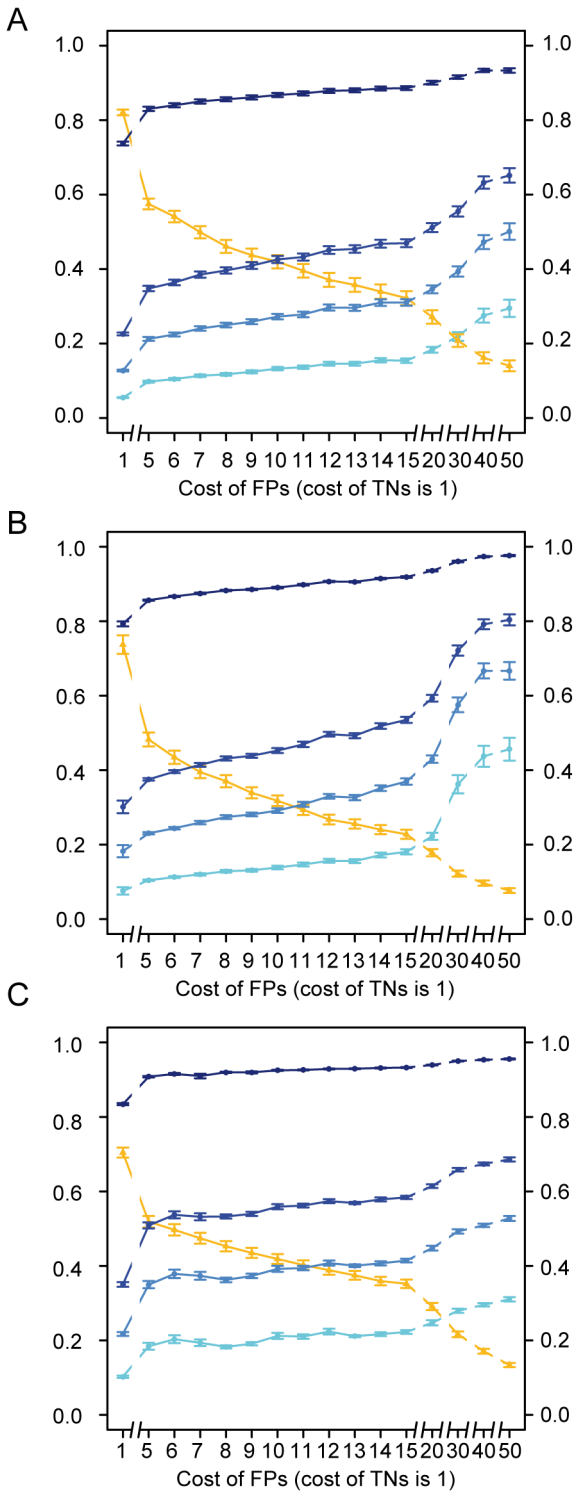

A
1.0
0.8
0.6
0.4
0.2
0.0
1 5 6 7 8 9 10 11 12 13 14 15 20 30 40 50
Cost of FPs (cost of TNs is 1)
B
1.0
0.8
0.6
0.4
0.2
0.0
1 5 6 7 8 9 10 11 12 13 14 15 20 30 40 50
Cost of FPs (cost of TNs is 1)
C
1.0
0.8
0.6
0.4
0.2
0.0
1 5 6 7 8 9 10 11 12 13 14 15 20 30 40 50
Cost of FPs (cost of TNs is 1)

Figure 5

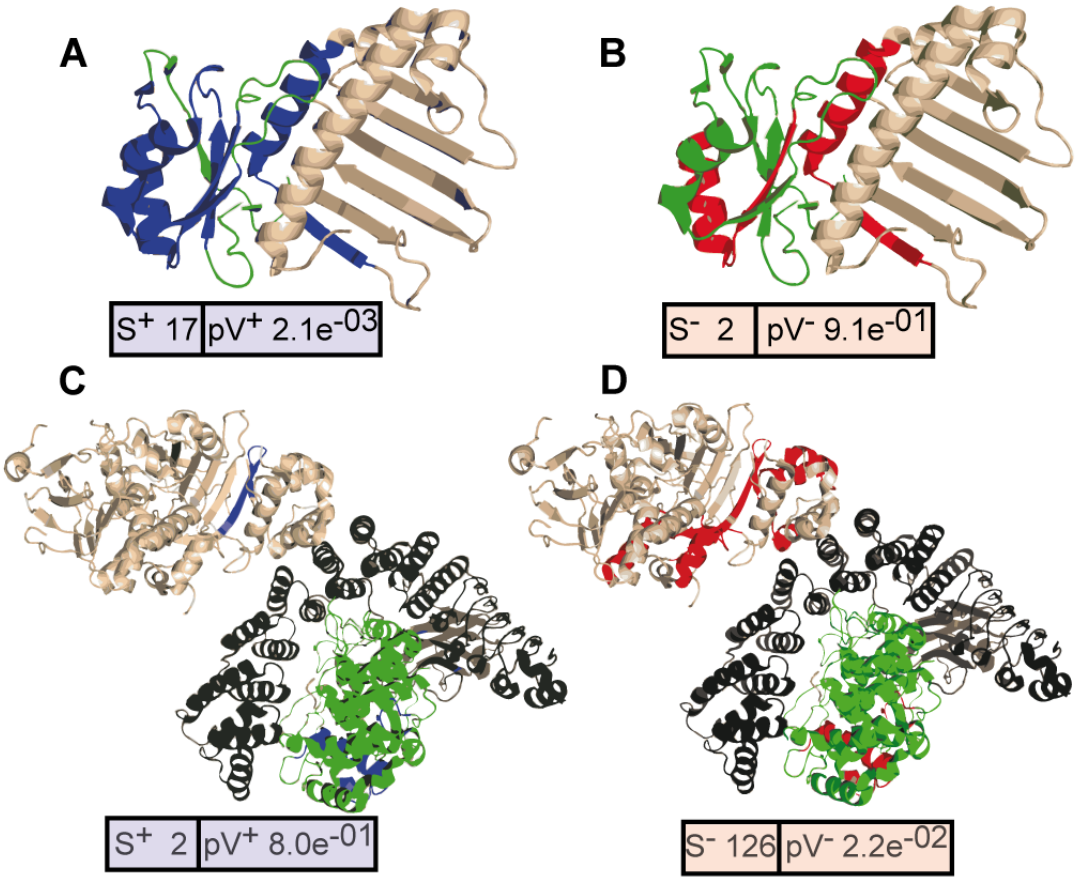


Figure 6

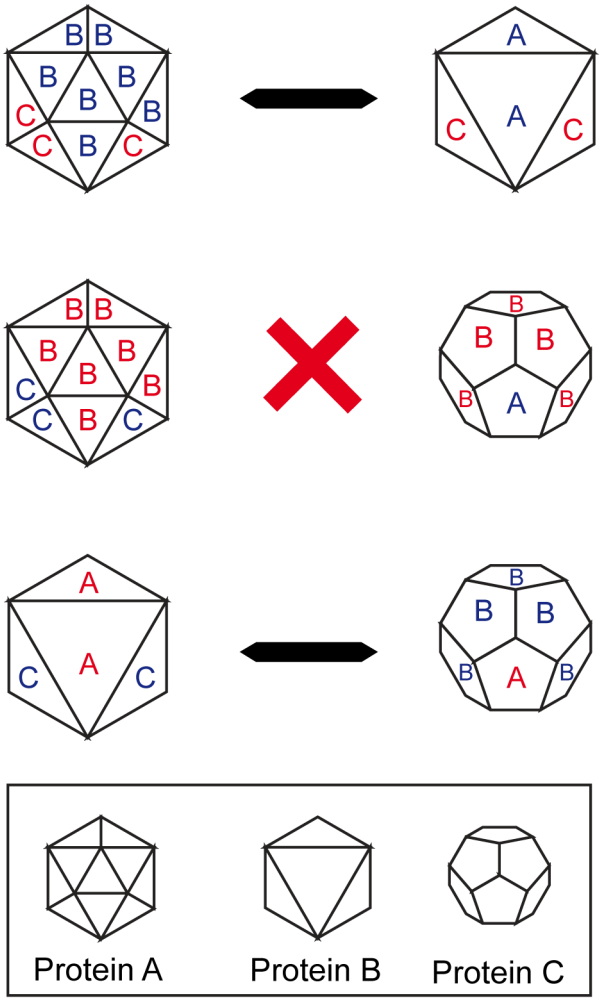

B B
B B
B
C B B
C B C
A
A
C C
B B
B B
B
C B B
C B C
B
B B
B A B
A
A
C C
B
B B
B A B
Protein A
Protein B
Protein C

Table 1

Table 1. AUCs of PPI classifiers.

	AUC								
Sign.	**pV^+**	**pV^-**	**LpVR**	**S^+**	**S^-**	**LSR**	**LO^+**	**LO^-**	**LOR**
{L}	0.67	0.79	0.82	0.59	0.82	0.85	-	-	-
$\{L_D\}$	0.67	0.78	0.82	0.60	0.81	0.86	-	-	-
{D}	0.61	0.70	0.86	0.56	0.73	0.89	0.66	0.62	0.93

Table 2

Table 2. Classification of PPIs and NIPs using LpVR and LSR.

Sign.	**TPR_{PPI}**	**TPR_{NIP}**	**PPV_{PPI}**	**PPV_{NIP}**
{L}	0.86	0.75	0.77	0.84
$\{L_D\}$	0.88	0.74	0.77	0.86
{D}	0.92	0.78	0.81	0.91

Table 3

Table 3. Performance of the prediction of PPIs.

	Unbalance-Ratio				
	1:1	**1:10**	**1:20**	**1:50**	**1:50**
Cost	**PPV**	**PPV**	**PPV**	**PPV**	**$TPR^{\dagger}$**
1:1	0.74 (0.06)	0.23 (0.04)	0.13 (0.03)	0.05 (0.01)	0.82 (0.01)
1:5	0.83 (0.07)	0.35 (0.09)	0.21 (0.06)	0.10 (0.03)	0.57 (0.17)
1:10	0.87 (0.07)	0.43 (0.11)	0.27 (0.09)	0.13 (0.05)	0.42 (0.21)
1:20	0.90 (0.06)	0.51 (0.14)	0.35 (0.13)	0.18 (0.09)	0.27 (0.21)
1:30	0.92 (0.05)	0.55 (0.17)	0.39 (0.16)	0.22 (0.14)	0.21 (0.20)
1:40	0.93 (0.05)	0.63 (0.20)	0.47 (0.22)	0.27 (0.22)	0.16 (0.18)
1:50	0.93 (0.07)	0.65 (0.23)	0.50 (0.26)	0.29 (0.27)	0.14 (0.17)

Table 4

Table 4. Comparison of PPI prediction methods.

Reference	Method	Input	Positive Database		Negative Database		%ID	Ratio	Evaluation					
			Training	Test	Training	Test			SEN	SPC	PPV	ACC	F1S	AUC
Jang et al. [44]	Correlation	Domains	DIP, Intact, MINT	iPfam	Random	Random	99%	1:1	0.82	0.83	-	-	0.87	-
Bjorkholm et al. [45]	Net. Weight	Domains	9 DIN	DOMINE	NO	NO	99%	-	-	-	0.53	-	-	-
Lee et al. [46]	MLE + Bayesian	Domains	Y2H&HPRD	iPfam, PQS	NO	NO	99%	-	0.33	-	-	-	-	-
Maetschke et al. [47]	Shortest Path	GO	STRING	BIND	Random	Random	99%	-	-	-	-	-	-	0.88
Shen et al. [48]	SVM	Sequence	HPRD	HPRD	Random	Random	-	1:1	0.84	-	-	0.83	-	-
iLoops	Correlation	Sequence	PES	PES	NES	NES	40%	1:1	0.82	0.69	0.74	0.76	0.77	0.84
iLoops	Correlation	Sequence	PES	PES	NES	NES	40%	1:50[†]	0.27	0.95	0.18	0.94	0.17	0.81

Supplementary Material for "Understanding protein-protein interactions using local structural features"

Index

***Table S4**. Classification of PPIs and NIPs using hyperbolic separating functions of pV^+ and S^+ (for positive signatures) and pV^- and S^- (for negative signatures).*

***Table S5**. Averaged PPV and TPR of the prediction of PPIs using $\{L_D\}$ interaction signatures and random forests classifiers*

***Table S6**. Averaged PPV and TPR of the prediction of PPIs using {D} interaction signatures and random forests classifiers.*

***Table S7.** Separating PPIs and NIPs with discriminant functions.*

***Table S8**. Classification of PPIs and NIPs using linear separating functions of LpVR and LSR.*

***Table S9**. Sizes of the Evaluation Sets (PES and NES).*

***Table S10**. Sizes of the subsets of pairs selected from the PES to test the random forest classifier.*

Supplementary Note 1: Details on PPI classifiers.

Let *(A,B)* be a pair of proteins. Let PS_A = *{ps_{A1}, …, ps_{Ai}}* and PS_B = *{ps_{B1}, …, ps_{Bj}}* be the sets of protein signatures of a certain type ({L}, {D}, or {L_D}), where ps_{Xn} represents a particular protein signature (X= *A* or *B)*. We define an *interaction signature* as a pair of protein signatures of the same type (one from protein A and the other from protein B). There were three different types of interaction signatures formed by pairs of {L}, {D}, and {L_D} protein signatures. They were labelled as {L}, {D}, and {L_D} interaction signatures, respectively. Formally, the set of interaction signatures is defined as: $IS_{(A,B)}$ = *{(ps_{A1}, ps_{B1}),(ps_{A1}, ps_{B2}),…, (ps_{Ai}, ps_{Bj})}*. Interestingly, even though an interaction signature *(ps_{Ai}, ps_{Bj})* could be observed in the PRS and the NRS sets, most interaction signatures in the PRS were not in the NRS. These observations suggested that positive and negative interaction signatures could be exploited to characterise PPIs and NIPs. Hence, we defined the *positive interaction signatures* ($IS_{(A,B)}^{+}$) and *negative interaction signatures* ($IS_{(A,B)}^{-}$) as the subsets of the interaction signatures in $IS_{(A,B)}$ that have non-zero entry in the frequency table derived from the PRS or the NRS respectively. Using these definitions, we define the following classifiers:

$$
\begin{aligned}
S^{+} &= \left| IS_{(A,B)^{+}} \right| \\
S^{-} &= \left| IS_{(A,B)^{-}} \right| \\
pV^{+} &= \min_{(ps_A, ps_B) \in IS_{(A,B)^{+}}} \left\{ pvalue\left(ps_A, ps_B\right) \right\} \\
pV^{-} &= \min_{(ps_A, ps_B) \in IS_{(A,B)^{-}}} \left\{ pvalue\left(ps_A, ps_B\right) \right\} \\
LO^{+} &= \max_{(ps_A, ps_B) \in IS_{(A,B)^{+}}} \left\{ \log odds\left(ps_A, ps_B\right) \right\} \quad \text{(SEq.1)} \\
LO^{-} &= \max_{(ps_A, ps_B) \in IS_{(A,B)^{-}}} \left\{ \log odds\left(ps_A, ps_B\right) \right\} \\
LSR &= \log 2\left(S^{+} / S^{-}\right) \\
LpVR &= -\log 2\left(pV^{+} / pV^{-}\right) \\
LOR &= LO^{+} - LO^{-}
\end{aligned}
$$

This is, we define pV^{+} and pV^{-} as the lowest p-value among those of the signatures in $IS_{(A,B)}^{+}$ and $IS_{(A,B)}^{-}$ respectively. Similarly, we denote the highest

log-odds value of positive and negative signatures in $IS_{(A,B)}^{+}$ and $IS_{(A,B)}^{-}$ as LO^{+} and LO^{-} respectively.

We emphasize that S^{+}, pV^{+}, and LO^{+} are meant to describe PPIs and are obtained from the frequency table built with the PRS. Conversely, S^{-}, pV^{-}, and LO^{-} are expected to explain NIPs and are derived from the frequency table built with the NRS. Finally, the ratio-based classifiers (*LSR*, *LpVR*, and *LOR*) require both frequency tables and are intended to classify PPIs.

Frequently, a protein interaction signature will be only included in the positive or the negative subset. In these cases, the ratio-based classifiers result in a zero-denominator division or a zero logarithm. To avoid such indeterminacies, the following pseudo-numbers are applied: the minimum number of signatures is $0+10^{-10}$; the worst (maximum) p-value is $1-10^{-10}$; and the best (minimum) p-value is 10^{-99}. For the computation of the potential PPI descriptors, any number beyond these limits is coerced into the described pseudo-numbers.

Supplementary Note 2: Details on the five-fold cross-validation

The goal of the five-fold cross-validation is to assess the ability of the potential PPI descriptors to classify the protein pairs in the test set as PPIs or NIPs. Protein pairs in the test set are classified as PPIs if the score for a potential PPI descriptor is above a threshold. On the contrary, protein pairs in the test set below this threshold are NIPs. Within this framework, a correctly classified PPI from the test set is considered a True Positive (TP) while a correctly classified NIP is a True Negative (TN). Conversely, a NIP in the test set reported as an interacting pair would be a False Positive (FP) and a PPI considered not to interact a False Negative (FN). These definitions apply to most of our potential PPI descriptors: S^{+}, pV^{+}, *LSR*, and *LpVR*. The other two descriptors, S^{-} and pV^{-}, are designed to classify in the opposite way, that is, to identify NIPs instead of PPIs. In these cases, NIPs and PPIs from the test set correctly classified are considered TPs and TNs respectively; PPIs predicted not to interact are FPs; and NIPs classified as interacting pairs are FNs. The True Positive Rate (TPR), False Positive Rate (FPR) and Positive Predictive Value (PPV) can be computed from these definitions (see Methods and equation 3 in the main text).

It has to be noted that the size of the PRS and the NRS differ. This is not a problem for the computation of TPR or FPR, since it involves the number of correctly and incorrectly classified protein pairs in only one of these sets. However, the PPV is a ratio between number of elements in the PRS and the NRS classified above a certain threshold; thus, its value depends on the relative sizes of these sets. The actual size of the interactome is still under discussion[1; 2; 3; 4; 5; 6]. Thus, determining the proper ratio of PRS and NRS sizes for the computation of PPV remains a challenge. Here, we chose to randomly sample ten times *n* elements of the larger set (where *n* is the size of the smallest set).

The five-fold cross-validation procedure involved two steps. First, protein pairs in the PRS or the NRS were split randomly into 5 groups. Then, four of the groups from the PRS and the NRS were selected as the *training set* and used to derive scores for positive and negative interaction signatures. The remaining group was used as *test set* to evaluate the prediction in the test set. The process was repeated 5 times.

To avoid sequence redundancy in the training process, we removed from the training set all protein pairs aligned with more than 99% of sequence identity with any protein pair in the test set. Then, to avoid biases by close homologs, we repeated the 5-fold cross validation removing from the training set all pairs aligned with more than 40% of sequence identity with any pair in the test set (see Supplementary Note 4).

For each prediction, we computed the averaged PPV values with the ten samples. With this procedure we have the same number of positive and negative scores and the probability of having a positive prediction by random is 0.5.

Supplementary Note 3: Non-redundant training sets.

To avoid homology redundancy, we trim the training sets by removing protein pairs with more than 40% sequence identity to protein pairs in the test set. Let (A,B) and (A',B') be two protein pairs; let seqID(A,A'), seqID(A,B'), seqID(B,A'), and seqID(B,B') be the percentage of sequence identity between the pairs (A,A'), (A,B'), (B,A'), and (B,B') respectively. We consider the pair (A,B) homologous to (A',B') if:

i. seqID(A,A') ≥ 40% and seqID(B,B') ≥ 40%, and/or

ii. seqID(A,B') ≥ 40% and seqID(B,A') ≥ 40

To obtain the non-redundant training sets, first we identify all homologous regions from proteins in the PRS and the NRS by BLASTing [7] them all against all. Then, a protein pair (X,Z) is removed from the training set if there exists a pair (X',Z') in the test set homologous to (X,Z).

Supplementary Note 4: Separating functions built from a combination of different PPI classifiers.

Our PPI classifiers are based either on the number of interaction signatures from a protein pair (S^+, S^-, *LSR*) or their best p-values (pV^+, pV^-, *LpVR*). Three logical combinations can be made using these classifiers: S^+ and pV^+ (positive combination), S^- and pV^- (negative combination), and *LSR* and *LpVR* (ratio combination). We consider the plane defined by the variables involved in each combination: x for the variable in (S^+, S^-, *LSR*) and y for the variable in (pV^+, pV^-, *LpVR*). A separating function in this plane is expected to separate PPIs from NIPs. Note that in the positive and ratio combinations, PPIs are expected to have a high number of signatures (or signatures ratio) and low best p-values (or p-value ratios), whereas in the negative combination the opposite holds. Hence, when representing S^+ as a function of $-\log(pV^+)$ or *LSR* as a function of *LpVR*, a good separating function should be placed above NIPs. On the contrary, if S^- is represented as a function of $-\log(pV^-)$, the separating function should be above PPIs. Supplementary Table S7 summarises the combinations used and what side of the separating function a correctly classified PPI or NIP should lie in.

The distribution of the PPI and NIP scores in a plane are separated differently for each combination of descriptors (see Figure 3 in the main text and supplementary Figures S2, S3, and S4); thus, the shape of the separating function has to differ as well. For the positive and negative combination, the separating function that discerns PPIs from NIPs should be a concave curve derived from a negatively sloped line. Let *A* and *B be* the y-intercept and the x-intercept points of such a line respectively, such that the line function is:

$$f(x) = A - \frac{A}{B}x \qquad \text{(SEq.2)}$$

The simplest polynomial function to describe the concave curve derived from this line is represented by:

$$f(x) = K_1 + \frac{K_2 x}{K_3 + K_4(x - K_5)} \quad \text{(SEq.3)}$$

where the following restrictions must be applied:

$$1.\ f(0) = A$$
$$2.\ K_4 = 0 \Leftrightarrow f(x) = A - \frac{A}{B}x \quad \text{(SEq.4)}$$
$$3.\ f(B) = 0$$

From restriction 1 we obtained that $K_1 = A$; to fulfil restriction 2 we choose $K_2 = -A$ and $K_3 = B$; and to fulfil restriction 3 we have $K_5 = B$. With these equalities, equation (S3) becomes:

$$f(x) = A - \frac{Ax}{B + K(x - B)} \quad \text{(SEq.5)}$$

For 0 < K < 1, this equation provides the desired curve. The larger the K, the more concave the curve is.

The separating function is optimised by testing several values for *A*, *B*, and *K*:

A = {*A'*/3, (2**A'*)/3, *A'*, (4**A'*)/3, (5**A'*)/3, 2**A'*}

B = {*B'*/10, *B'*/9, *B'*/8, *B'*/7, *B'*/6, *B'*/5, *B'*/4, *B'*/3, *B'*/2, *B'*}

K = {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99}

where: *A'* is the maximal S^+ for the positive combination or the maximal S^- for the negative combination; *B'* is the maximal $-\log(pV^+)$ or $-\log(pV^-)$ for the positive and negative combinations respectively. For a protein pair with scores (x,y), y being larger or smaller than f(x) determines in which side of the curve the point (x,y) lies.

An inspection of the distribution of the PPI and NIP scores in the ratio combination suggests that a negatively sloped line discriminates them appropriately. Such a line can be defined as:

$$f(x) = -mx + n \quad \text{(SEq.6)}$$

where -*m* is the slope of the line and n the y-intercept point. If α is the angle of this line with the x-axis, then:

$$m = \tan(\alpha) \quad \text{(SEq.7)}$$

To optimise the separating function we use different values of α increasing from 0º to 90º in a 2.5º step. We also test different *n* values: we use a zero-centred distribution of 21 equally spread values of *n* in the interval [-a,a] with a=max(|*v*|, |*V*|)/3 where *V* and *v* are the maximum and minimum *LpVR* values in the five-fold cross-validation respectively.

We considered as the optimum separating function the one that maximised the following optimisation criterion (OC):

$$OC = \frac{\left(TPR_{PPI} * FPR_{PPI} - FPR_{NIP} * TPR_{NIP}\right)}{\sqrt{\left(TPR_{PPI} + FPR_{PPI}\right)\left(TPR_{NIP} + FPR_{NIP}\right)\left(TPR_{PPI} + FPR_{NIP}\right)\left(TPR_{NIP} + FPR_{PPI}\right)}} \quad \text{(SEq.8)}$$

Supplementary Note 5: Analysis of positive interaction classifiers.

First, we assessed if the p-value (equation (1) from the main text) was equivalent to the log-odds score[8] in equation (2) (from the main text). Similarly to the work of Sprinzak and Margalit[8], we neglected sequence similarity and predicted PPIs using pV^+ and LO^+ using positive domain interaction signatures {D}. We compared the ROC curves (see Figures 2C and 2D in the main text), their AUCs (supplementary Table S2) and the associated estimation of the error of AUC (see supplementary Table S3). The AUC using pV^+ 0.64, while the AUC using LO^+ was 0.68. Therefore, we considered that the p-value measure was comparable to the log-odds score used by Sprinzak and Margalit[8] to predict PPIs.

Second, we extended the analysis to {L} and {L_D} positive interaction signatures for the prediction of PPIs. Interestingly, the AUC obtained with the pV^+ classifier when the similarity between pairs of proteins was not removed (99% ID) was larger than 0.85 for both types of interaction signatures. However, after removing pairs with more than 40% sequence identity in the PRS the results were similar to those obtained with domain interaction signatures (see Table 1 in the main text and supplementary Table S3). This suggested that the predictions based on loops encompassed a bias due to sequence similarities. Therefore, it could be argued that the capacity of the pV^+ classifier to identify new PPIs using {L} or {L_D} positive interaction signatures was limited (see average ROC curves in Figures 2A and 2B in the

main text), or at least as limited as predicting with domains. This also suggested that interaction signatures based on loops recapitulated the information of domains when used for predicting PPIs. In conclusion, the average AUC of the prediction using the pV^+ classifier was not satisfactory to discern between new PPIs and NIPs.

Next, we applied the S^+ classifier. The analyses of ROC curves showed similar results as those obtained with pV^+ (Figures 2E, 2F, and 2G in the main text). The average AUCs obtained in the 40%ID set were still similar to the results of a random classifier (see Table 1 in the main text), but with low associated errors (see supplementary Table S3).

Although the success of pV^+ and S^+ classifiers to discriminate between PPIs or NIPs was poor, we also observed that NIPs tended to show either a high p-value or a low number of signatures (see supplementary Figure S2). Exploiting this observation, we designed hyperbolic separating functions to better discern between PPIs and NIPs (see Supplementary Note 4). It has to be noted that this was a classification of all PPIs and NIPs, instead of 1/5 of the 5-fold approach. These hyperbolic separating functions correctly identified around 55% of PPIs with less than 32% of NIPs (supplementary Table S4) when using loop interaction signatures ({L} and $\{L_D\}$) and proteins pairs with less than 40%ID. Interestingly, the percentage of correct predictions improved when the test included pairs with similar sequences (99%ID), proving the bias produced by homology. This improvement was mostly noticed when using loop signatures instead of domain signatures.

Supplementary Note 6: Analysis of negative interaction classifiers.

In the previous section we studied the correlation between pairs of interacting proteins and specific protein features involved in the interaction. We investigated how to use this correlation to classify and putatively predict PPIs. Analogously, some kind of correlation was suspected for structural features hindering the interaction between a pair of proteins. Therefore, we derived the classifiers pV^- and S^- with the objective of predicting pairs of non-interacting proteins, and we expected to identify NIPs with good scores. Additionally, we also applied the log-odds score (LO^-), although this could not be compared with the original work of Sprinzak and Margalit[8]. The analysis of the average

ROC curves showed the success of the prediction when using the pV^- classifier for all types of signatures (Figures 2A, 2B, and 2C in the main text, supplementary Table S2), while the LO^- classifier was not much better for predicting NIPs than it was for predicting PPIs. It was noteworthy that removing pairs of similar sequences (>40%ID) did not affect the ROC curves. These trends were also observed in the averaged AUCs (Table 1 in the main text), with values larger than 0.7. Not surprisingly, a bias by sequence homology cannot take place between pairs of non-interacting proteins. We argue that homologs of a pair of interacting proteins have high probabilities to preserve the interaction, while homologs of non-interacting pairs have the same chances to interact (or not) as any other pair.

As in the analysis of positive signatures, we also studied the number of negative signatures with similar justification: a protein pair with a large number of non-interaction signatures would rarely interact. The S^- classifiers of loop interaction signatures {L} and {L_D} were applied using a five-fold approach (Figures 2E and 2F in the main text show the average ROC curves; averaged AUCs are summarized in Table 1 in the main text and in supplementary Table S2). The results were similar to the previous results obtained with the pV^- classifier and the same trend was observed when using domain signatures (see Table 1 and Figure 2G in the main text, and supplementary Table S2).

Next, we used hyperbolic separating functions of pV^- and S^- as in the previous section to separate PPIs and NIPs (see Supplementary Note 4). Similarly to what we observed with the classifiers of positive signatures, NIPs tended to have either large number of negative signatures (S^-) or low pV^-, while PPIs had pV^- values close to 1 and low S^- (see supplementary Figure S3). For instance, using {L} negative interaction signatures and removing homology from the training sets, the separating function correctly identified more than 50% of NIPs while only misclassifying less than 3% (see supplementary Table 4).

Supplementary Note 7: Extension of the analysis of log-ratio classifiers: *LpVR*, *LSR*, and *LOR*

The results presented in the main text correspond to the classifiers trained with non-homologous protein pairs (maximum sequence identity = 40%). In

order to allow comparison of our results with previous works°, we include here the results of the analysis of the classifiers trained with homologous protein pairs for comparative purposes.

The average ROC curves for *LpVR* showed that this classifier obtained better results than either of pV^+ or pV^- classifiers trained with homologous protein pairs (see supplementary Figures S1A, S1B, and S1C, and supplementary Table S8). Regarding the *LSR* classifier, the average ROC curves and AUCs for loop and domain signatures showed the same trends as for *LpVR* (see Table and Figures 2E, 2F and 2G in the main text). Predictions using this classifier were slightly better than using *LpVR*, although for all classifiers the errors associated with the AUCs were very small (see supplementary table S3). Moreover, the obtained averaged AUCs were larger than 0.95 for all types of signatures ({L}, $\{L_D\}$, and {D}).

Combining both classifiers, we designed a linear separating function to discern between PPIs and NIPs (see supplementary Note 4). We separated above 95% of PPIs with less than 15% misclassified NIPs using {L} or $\{L_D\}$ signatures, or up to 20% using {D} signatures (see supplementary Table 8, supplementary Figure S4). The separation was similarly good to distinguish NIPs from PPIs. Thus, the values of PPV_{PPI}, calculated as the ratio of true PPIs among the total pairs classified as putative interactions, and NPV_{PPI} (i.e. PPV_{NPI}), calculated as the ratio of true NIPs over the pairs classified as non-interacting, were close. For instance, for {L} signatures, the separating function achieved 0.89 PPV_{PPI} and 0.97 PPV_{NPI} (see supplementary Table 8).

Supplementary Note 8: Details of the additional evaluation

We designed an additional validation of our classifiers to test their predictive power. New sets of PPIs (Positive Evaluation Set, PES) and NIPs (Negative Evaluation Set, NES) were obtained as described in Methods section in the main text. Each of these sets was sub-divided into three groups: one for building frequency matrices, a second to train a random forest classifier[9], and a third to test this classifier. Supplementary Table S9 summarises the sizes of the sets according to the type of signature used.

We used the WEKA package[10] to apply this strategy. We considered the previously described classifiers (pV^+, pV^-, S^+, S^-, *LpVR*, and *LSR*), and we

included other parameters that could describe the distribution of p-values of the interaction signatures. First, we considered the 10 best (lower) p-values of positive and negative signatures. Next, we included as training parameters some values of the distribution of p-values of positive and negative interaction signatures, such as the minimum, the maximum, the mean, and the first, second, and third quartiles. Finally we considered the absolute number of residues covered by positive or negative interaction signatures as well as their relative coverage, this is the number of residues covered over the sum of the sequence length of the protein pair.

We repeated the training several times (training replicas) including in each repetition a different combination of protein pairs from the NES and the PES. We used 55, 40, and 100 different training sets for {L}, {L_D}, and {D} signatures respectively. To simulate natural unbalance between PPIs and NIPs (NIPs are much more frequent than PPIs)[6; 11] we tested the classifier using all the NIPs available (see *NES Evaluation* row in supplementary Table S10) and different amounts of PPIs in the PES, obtaining evaluation sets with the following proportions between PPIs and NIPs: 1:10, 1:20 and 1:50. We refer to this proportion as *unbalance ratio* (UR), and supplementary Table S11 details the number of PPIs used in each UR test (the corresponding number of NIPs is trivial).We performed an additional test to simulate an UR of 1:1 by means of a 10-fold cross-validation of the training data.

We have to note that the set of co-localized proteins represents around 5% of all possible protein pairs in a human cell. Thus, the 1:50 ratio simulates the naturally occurring unbalance between PPIs and NIPs in the human proteome (which has been estimated to be about 1 PPI for 1000 NIPs[6;11]). Consequently, these ratios represent different levels of expertise of a user that applies the prediction method (from an ideal expert, who is able to select pairs with a 50% probability to have an interaction, to a non-expert, who almost randomly selects any pair of proteins of the interactome with ordinary good judgement).

The results with a 1:1 ratio of PPIs versus NIPs were similar to those obtained for the classification with the sets PRS and NRS. Nevertheless, testing the classifier with an increasing proportion of NIPs over PPIs produced the decrease of PPV. Therefore, increasing the relative cost of accepting

incorrectly predicted PPIs showed a remarkable improvement that we think can be very interesting for an experimentalist. In the worst scenario of the 1:50 ratio, the increase of relative-cost from 1:1 to 1:20 increased the PPV from 0.05 to 0.18 at the expense of decreasing the TPR from 0.82 to 0.27. Table 3 in the main text summarizes the PPV and TPR when using {L} interactions signatures at different URs and relative costs; supplementary Table S5 shows the results for $\{L_D\}$ interaction signatures, and supplementary Table S6 for {D} interaction signatures. Figure 6 in the main text shows the averaged PPV (and the associated error) for different ratios and relative-costs to penalize false positives when predicting PPIs using {L}, $\{L_D\}$ and {D} interaction signatures.

We wish to note that our tests tried to simulate the problem of experimental biologists who want to use prediction methods of PPIs: first, because they would rarely test a pair of proteins known to be compartmentalised apart in the cell; second, because they usually have some intuition on the putative protein partners; and third, because they may wish to select the best conditions either to obtain the largest number of real interactions (highest recall) or to ensure with few experiments the minimum number of failures (highest PPV).

Supplementary Note 9: Prediction of the interactions in the human exosome

One particular case to distinguish between NIPs and PPIs is the exosome complex. Predicting the interactions between homolog proteins in the same structural complex can be complicated. The human RNA core exosome is composed by one ring of six proteins (the complex structure is found in the PDB with code 2nn6[12]). The six proteins are split in two groups of three close homologs, while the sequence similarity between any pair of proteins of each group is low (remote homologs), and all proteins of the core exosome have the same fold. As a first approach we assumed that all proteins with the same fold interacted, but we failed to form the exosome complex (the specificity of such prediction is null). Then, in a second approach, we tried to predict PPIs using the linear separating classifier with loop interaction signatures {L}. This example showed the difficulty to distinguish the interacting and non-interacting pairs (these being too close to the separating line). Nevertheless, this second

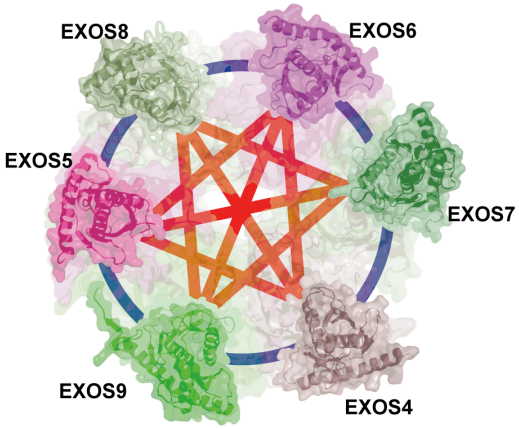
approach improved the previous prediction based on fold similarity (the specificity increased to 0.89) (see supplementary Figure S1).

References

1. Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., Rual, J. F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P. & Vidal, M. (2009). Literature-curated protein interaction datasets. *Nat Methods* **6**, 39-46.
2. Hart, G. T., Ramani, A. K. & Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol* **7**, 120.
3. Kelly, W. P. & Stumpf, M. P. (2012). Assessing coverage of protein interaction data using capture-recapture models. *Bull Math Biol* **74**, 356-74.
4. Sambourg, L. & Thierry-Mieg, N. (2010). New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics* **11**, 605-15.
5. Tompa, P. & Rose, G. D. (2011). The Levinthal paradox of the interactome. *Protein Sci* **20**, 2074-9.
6. Venkatesan, K., Rual, J. F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K. I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A. S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabasi, A. L. & Vidal, M. (2009). An empirical framework for binary interactome mapping. *Nat Methods* **6**, 83-90.
7. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology* **215**, 403-10.
8. Sprinzak, E. & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of molecular biology* **311**, 681-92.
9. Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5-33.
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **11**.
11. Yu, C. Y., Chou, L. C. & Chang, D. T. (2010). Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics* **11**, 167.
12. Liu, Q., Greimann, J. C. & Lima, C. D. (2006). Reconstitution, activities, and structure of the eukaryotic RNA exosome. *Cell* **127**, 1223-37.

Supplementary Figures

Figure S1. Representation of the analysis in the eukaryotic RNA exosome.



Interaction	S^+	S^-	pV^+	pV^-	Prediction
EXOS4-EXOS7	15	498	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS4-EXOS9	15	498	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS5-EXOS8	13	158	$2.09e^{-2}$	$7.39e^{-3}$	✓
EXOS5-EXOS9	08	136	$5.29e^{-2}$	$7.39e^{-3}$	✓
EXOS6-EXOS7	04	178	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS6-EXOS8	06	216	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS4-EXOS5	11	077	$1.39e^{-4}$	$7.39e^{-3}$	✗
EXOS4-EXOS6	05	102	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS4-EXOS8	21	587	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS5-EXOS6	02	026	$1.79e^{-1}$	$1.49e^{-2}$	✗
EXOS5-EXOS7	08	136	$5.29e^{-2}$	$7.39e^{-3}$	✓
EXOS6-EXOS9	06	187	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS7-EXOS8	16	525	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS7-EXOS9	09	342	$8.99e^{-4}$	$7.39e^{-3}$	✗
EXOS8-EXOS9	16	525	$8.99e^{-4}$	$7.39e^{-3}$	✗

iLoops

PREDICT \ REAL	+	-	
+	TP 2	FP 1	P' 3
-	FN 4	TN 8	N' 12
	P 6	N 9	15 / 15

Domains

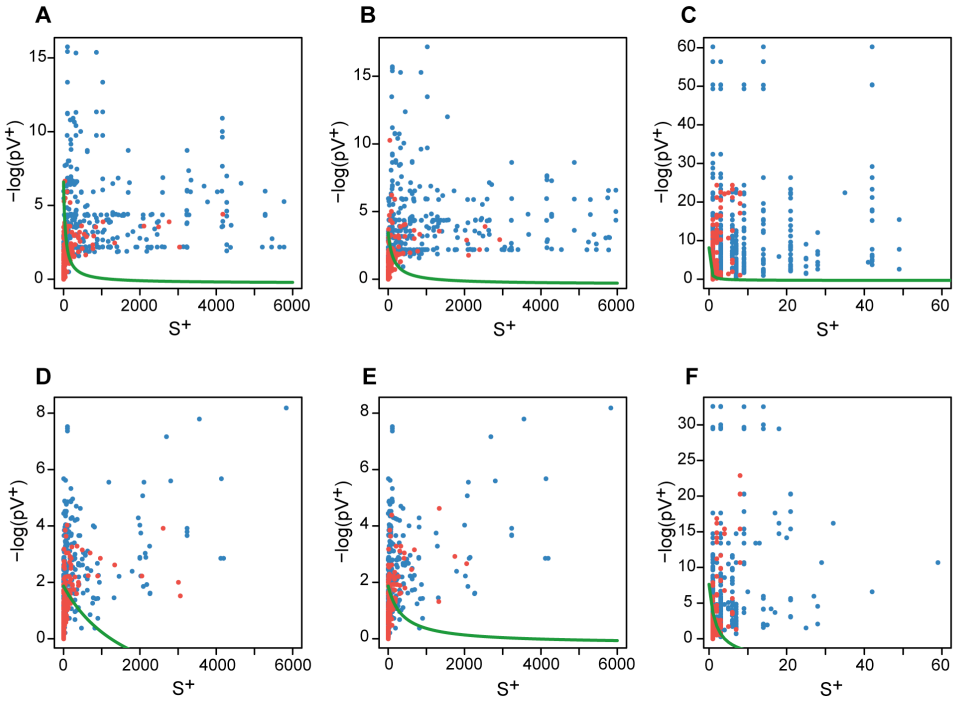
PREDICT \ REAL	+	-	
+	TP 6	FP 9	P' 15
-	FN 0	TN 0	N' 0
	P 6	N 9	15 / 15

0.66	PPV	0.40
0.89	1-FPR	0.00
0.33	TPR	1.00
1.00	App	1.00
0.27	MCC	0.00

The top left corner represents the interacting relations between the 6 proteins of the exosome ring. Blue lines represent PPIs while red lines represent NIPs. The left table represents the analysis for each putative interacting pair. Those with blue background represent the actual PPIs while those with red background represent NIPs. For each line, the values of S^+, S^-, pV^+ and pV^- classifiers are given. Predictions according to the random forest classifier are in the last column. The blue tick represents a PPI prediction and the cross in red a NIP prediction. Contingency tables for the analysis with iLoops and with domains are displayed in the left. The statistical analyses of both predictions are compared in the bottom left corner by means of the PPV, the specificity

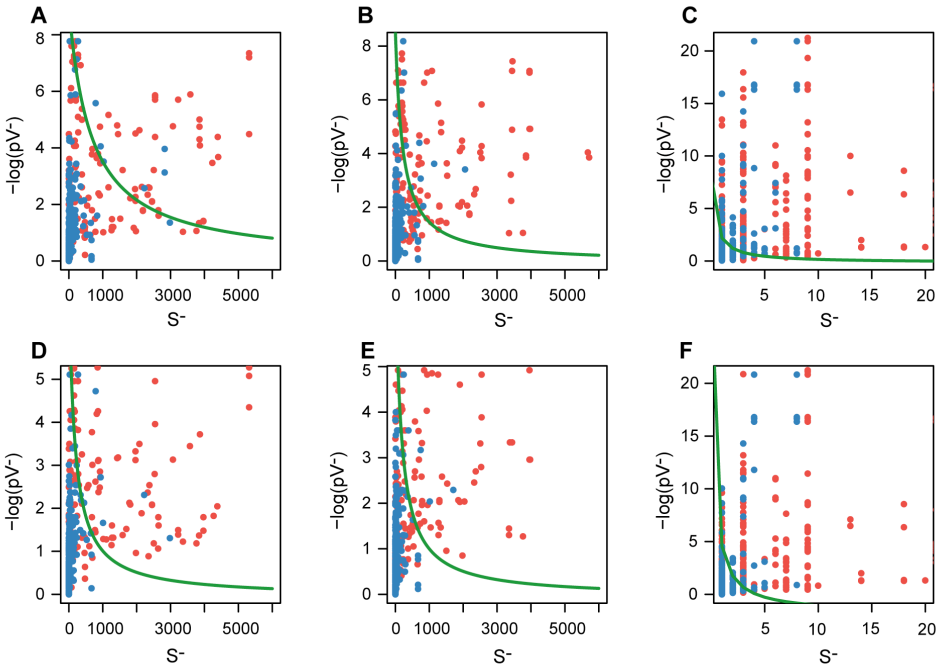
(1-FPR), the TPR, the applicability (App) and the Mathews Correlation Coefficient (MCC). The analysis shows that the iLoops can identify some correct relationships. Assuming that all proteins with the same fold interact yields null specificity and 40% PPV, while iLoops prediction using random forest and 1:1 relative cost produces 66%PPV and 89% specificity.

Figure S2. Separating functions to discern between PPIs and NIPs using pV^+ and S^+ classifiers.



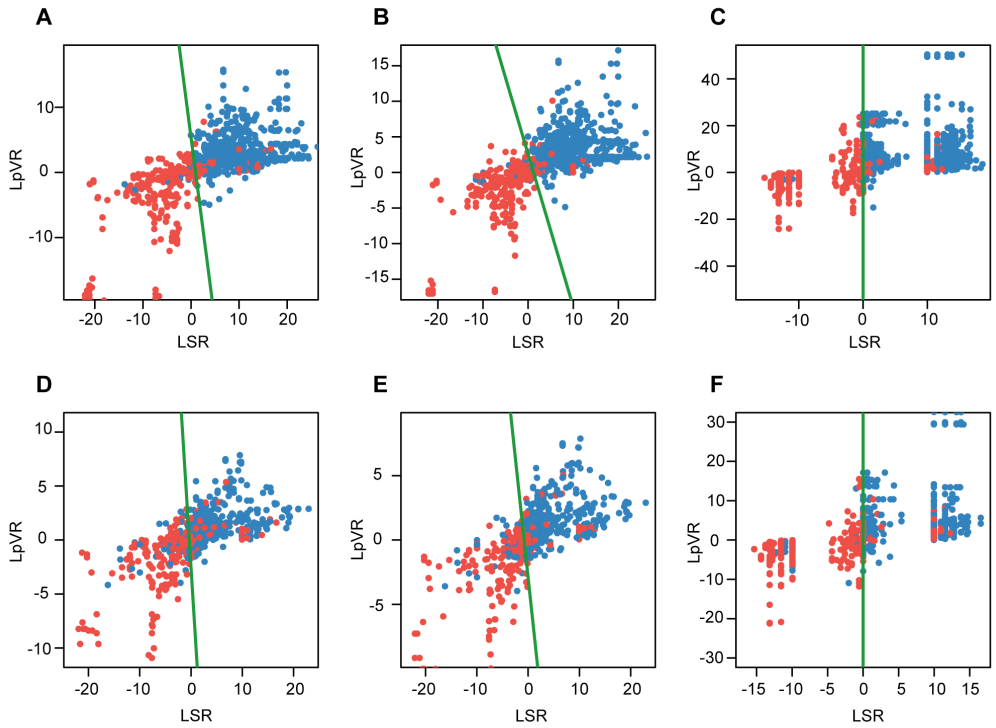
Distribution of pV^+ versus S^+ of PPIs (blue) and NIPs (red) are shown for {L} (A and D), {L_D} (B and E), and {D} (C and F) interaction signatures. Upper panels (A-C) show the results obtained with homology redundant pairs in the training sets, while lower panels (D-F) correspond to predictions made removing such homologous pairs from the training sets.

Figure S3. Separating functions to discern between PPIs and NIPs using pV^- and S^- classifiers.



Distribution of pV^+ versus S^+ of PPIs (blue) and NIPs (red) are shown for {L} (A and D), {L_D} (B and E), and {D} (C and F) interaction signatures. Upper panels (A-C) show the results obtained with homology redundant pairs in the training sets, while lower panels (D-F) correspond to predictions made removing such homologous pairs from the training sets.

Figure S4. Linear separating functions separating PPIs and NIPs.



Distribution of *LpVR* versus *LSR* of PPIs (blue) and NIPs (red) are shown for interaction signatures {L} (panels A and D), {L_D} (panels B and E), and {D} (panels C and F). The plots show the distribution of protein pairs with sequence identity smaller than 99% (panels A-C) and protein pairs with less than 40% sequence identity (panels D-F). Optimised line separating PPIs from NIPs are depicted in green.

Supplementary Tables

Table S1. Number of protein pairs, protein signatures, and interaction signatures.

Number of protein pairs, protein signatures and number of interaction signatures in the PRS and the NRS for different types of signatures (TS).

	#protein pairs		**#protein signatures**		**#interaction signatures**	
TS	**PRS**	**NRS**	**PRS**	**NRS**	**PRS**	**NRS**
{D}	2821	699	1522	998	7546	3904
{L}	632	309	89120	96142	16584271	10847908
$\{L_D\}$	632	288	60575	55546	9411279	3692270

Table S2. AUCs of PPI classifiers.

The AUCs for different classifiers trained containing homologous protein pairs (99% maximum sequence identity) are given in columns 2 to 10. First column shows the type of signature used in the classifier (Sign.).

	AUC								
Sign.	**pV^+**	**pV^-**	**LpVR**	**S^+**	**S^-**	**LSR**	**LO^+**	**LO^-**	**LOR**
{L}	0.86	0.83	0.93	0.86	0.84	0.96	-	-	-
$\{L_D\}$	0.89	0.83	0.93	0.86	0.84	0.96	-	-	-
{D}	0.63	0.71	0.87	0.58	0.74	0.95	0.68	0.61	0.93

Table S3. Error associated to AUCs of PPI classifiers.

Errors associated to the AUCs for different classifiers are given according to the maximum sequence identity (Max ID) allowed in the training set and the type of signature used (Sign.) in the classifier. These numbers correspond to the difference between the extreme values of the AUCs obtained at each round of the five-fold cross-validation and the AUC of the mean ROC curve.

		±AUC associated error								
Max ID	**Sign.**	**pV^+**	**pV^-**	**LpVR**	**S^+**	**S^-**	**LSR**	**LO^+**	**LO^-**	**LOR**
99 %	{L}	0.044	0.062	0.025	0.066	0.040	0.038	-	-	-
99 %	$\{L_D\}$	0.050	0.041	0.041	0.038	0.040	0.023	-	-	-
99 %	{D}	0.063	0.024	0.025	0.080	0.024	0.013	0.057	0.023	0.005
40 %	{L}	0.089	0.069	0.061	0.084	0.051	0.075	-	-	-
40 %	$\{L_D\}$	0.086	0.082	0.053	0.075	0.059	0.038	-	-	-
40 %	{D}	0.030	0.049	0.041	0.053	0.036	0.023	0.065	0.032	0.015

Table S4. Classification of PPIs and NIPs using hyperbolic separating functions of pV^+ and S^+ (for positive signatures) and pV^- and S^- (for negative signatures).

Coverage of PPIs (TPR_{PPI}), coverage of NIPs (TPR_{NIP}), precision of the classification of interactions (PPV_{PPI}), and precision of the classification of non-interactions (PPV_{NPI}) are shown in columns 4, 5, and 6 (see methods section). Columns 1-2 indicate the type of interaction signatures used for the classification (column 2) and if the classifiers were aimed at classify PPIs (positive interaction signatures) or NIPs (negative interaction signatures). Column 3 indicates the maximum sequence identity (Max ID) allowed between the training set and the test set.

+/- Sign.	Sign.	Max ID	TPR_{PPI}	TPR_{NIP}	PPV_{PPI}	PPV_{NIP}
Positive	{L}	99 %	0.86	0.77	0.91	0.71
Positive	$\{L_D\}$	99 %	0.89	0.78	0.92	0.70
Positive	{D}	99 %	0.92	0.31	0.62	0.63
Positive	{L}	40 %	0.56	0.71	0.75	0.54
Positive	$\{L_D\}$	40 %	0.55	0.68	0.74	0.53
Positive	{D}	40 %	0.56	0.69	0.59	0.59
Negative	{L}	99 %	0.97	0.46	0.64	0.95
Negative	$\{L_D\}$	99 %	0.97	0.55	0.68	0.95
Negative	{D}	99 %	0.74	0.74	0.74	0.74
Negative	{L}	40 %	0.97	0.51	0.66	0.95
Negative	$\{L_D\}$	40 %	0.98	0.44	0.63	0.96
Negative	{D}	40 %	0.82	0.68	0.72	0.79

Table S5. Averaged PPV and TPR of the prediction of PPIs using $\{L_D\}$ interaction signatures and random forests classifiers (standard deviations are shown between parenthesis). Columns 2-6 indicate the results for different unbalanced ratios of PPIs versus NIPs and the first column indicates the relative-cost of false-positives versus false-negatives applied in the random-forest classifier.

	Unbalance Ratio				
	1:1	1:10	1:20	1:50	1:50
Cost	PPV (sd)	PPV (sd)	PPV (sd)	PPV (sd)	TPR (sd)
1:1	0.79 (0.07)	0.30 (0.17)	0.18 (0.16)	0.08 (0.10)	0.74 (0.25)
1:5	0.86 (0.01)	0.37 (0.03)	0.23 (0.03)	0.10 (0.02)	0.48 (0.19)
1:10	0.89 (0.02)	0.45 (0.06)	0.29 (0.06)	0.14 (0.05)	0.32 (0.14)
1:20	0.93 (0.02)	0.59 (0.09)	0.43 (0.10)	0.22 (0.09)	0.18 (0.10)
1:30	0.96 (0.02)	0.72 (0.13)	0.57 (0.19)	0.36 (0.23)	0.12 (0.08)
1:40	0.97 (0.02)	0.79 (0.13)	0.67 (0.19)	0.44 (0.26)	0.10 (0.07)
1:50	0.98 (0.02)	0.80 (0.14)	0.67 (0.22)	0.47 (0.28)	0.08 (0.06)

Table S6. Averaged PPV and TPR of the prediction of PPIs using {D} interaction signatures and random forests classifiers (standard deviations are shown between parenthesis). Columns 2-6 indicate the results for different unbalanced ratios of PPIs versus NIPs and the first column indicates the relative-cost of false-positives versus false-negatives applied in the random-forest classifier.

	Unbalance Ratio				
	1:1	**1:10**	**1:20**	**1:50**	**1:50**
Cost	**PPV (sd)**	**PPV (sd)**	**PPV (sd)**	**PPV (sd)**	**TPR (sd)**
1:1	0.83 (0.05)	0.35 (0.10)	0.22 (0.08)	0.10 (0.05)	0.70 (0.22)
1:5	0.91 (0.03)	0.51 (0.13)	0.35 (0.15)	0.18 (0.14)	0.52 (0.25)
1:10	0.92 (0.02)	0.56 (0.10)	0.39 (0.12)	0.21 (0.12)	0.42 (0.21)
1:20	0.94 (0.02)	0.61 (0.07)	0.49 (0.10)	0.25 (0.10)	0.29 (0.16)
1:30	0.95 (0.02)	0.66 (0.07)	0.49 (0.09)	0.28 (0.08)	0.22 (0.12)
1:40	0.95 (0.02)	0.67 (0.06)	0.51 (0.07)	0.29 (0.06)	0.17 (0.10)

Table S7. Separating PPIs and NIPs with functions.

The first and second columns show the combination of input descriptors used in the separating function. Third and fourth columns indicate the positions of correctly classified PPIs and NIPs in relation to the separating line.

Combination	**Descriptors**	**PPI**	**NPI**
Positive combination	(S^+,pV^+)	Above the line	Below the line
Negative combination	(S^-,pV^-)	Below the line	Above the line
Ratio combination	(*LSR*,*LpVR*)	Above the line	Below the line

Table S8. Classification of PPIs and NIPs using linear separating functions of LpVR and LSR.

Ratios of correctly classified PPIs (TPR_{PPI}), NIPs (TPR_{NIP}), and positive predictive values for the classifcation of interactions (PPV_{PPI}) and non-interactions (PPV_{NIP}) using linear separating functions of *LpVR* and *LSR*. The results shown correspond to classifiers trained with homologous protein pairs (99% maximum sequence identity). The first column indicates the interaction signature applied for the classification.

Sign.	**TPR_{PPI}**	**TPR_{NIP}**	**PPV_{PPI}**	**PPV_{NIP}**
{L}	0.97	0.88	0.89	0.97
{L_D}	0.94	0.87	0.87	0.96
{D}	0.99	0.80	0.83	0.99

Table S9. Sizes of the Evaluation Sets (PES and NES).

The PES and NES have different sizes for each type of signature under study ({L}, {L_D} and {D}, shown in columns). The PES and NES were split in six training and evaluation subsets: two to obtain interaction signatures (PES signatures and NES signatures), two for training a random forest classifier (PES training and NES training), and two to test the prediction (PES evaluation and NES evaluation).

Signature	**{L}**	**{L_D}**	**{D}**
PES Signatures	1000	1000	4000
PES Training	500	500	500
PES Evaluation	5241	4277	3114
NES Signatures	1000	1000	4000
NES Training	500	500	500
NES Evaluation	1861	1478	14200

Table S10. Sizes of the subsets of pairs selected from the PES to test the random forest classifier.

The unbalanced ratio of PPIs and NIPs is shown in the first column. Columns 2-4 show the number of pairs randomly selected from the PES-Evaluation subset for testing interaction signatures {L}, {L}, and {D}. The size of the NES-Evaluation subset is trivial.

	Type of signature		
UR	**{L}**	**{L_D}**	**{D}**
1:10	185	147	1420
1:20	93	74	710
1:50	37	29	284

3.2.2 iLoops: A protein-protein interaction prediction server based on local structural features

iLoops web server manuscript:

Planas-Iglesias, J., Bonet, J., Marín-López, M. A. & Oliva, B. (2012). iLoops: A protein-protein interaction prediction server based on local structural features *Submitted to Bioinformatics*.

iLoops: A protein-protein interaction prediction server based on local structural features.

Joan Planas-Iglesias[1,†], Manuel A. Marin-Lopez[1,†], Jaume Bonet[1,†] and Baldo Oliva[1,*]

[1] Structural Bioinformatics Lab (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Research Park of Biomedicine (PRBB), Barcelona, Catalonia, 08950, Spain

† Authors contributed equally to this work
* To whom correspondence should be addressed. Tel: (+34) 933 160 509; Fax: (+34) 933 160 550; Email: baldo.oliva@upf.edu

ABSTRACT

Protein-protein interactions play a critical role in many biological processes. Despite that, the number of servers that provide an easy and comprehensive method to predict them is still limited. The iLoops server predicts if a pair of proteins can interact by exploring the relationships of its structural features (loops and domains). As a source of loop features the server uses ArchDB's hierarchy of loops and SCOP classification of domains. The input of the server are the sequences of the query proteins to be tested. Loops and domains are assigned to the query proteins by sequence similarity. The prediction algorithm uses information from known protein-protein interactions and confirmed non-interacting proteins. Known interactions were extracted from several databases of yeast-two-hybrid experiments. The Negatome database was used to obtain pairs of non-interacting proteins. Pairs of structural features (formed by loops or domains) were classified according to its likelihood to favour or disfavour a protein-protein interaction. The server uses the relationship between the protein-features assigned to a pair of query proteins to predict their interaction. The iLoops server is freely accessible at http://sbi.imim.es/iLoops.php

INTRODUCTION

Interactions between proteins mediate almost all the processes in a living cell. Thus, the discovery of new protein-protein interactions (PPI) is key to understand the complexity of the biological systems. Due to the biological relevance of PPIs, several experimental methods have been developed to identify new PPIs. Yeast two hybrid (Y2H) and tandem affinity purification (TAP) are among the most used methods for high throughput identification of new interacting proteins (1). However, these methods are still economically and timely costly, they lack reproducibility and yield a high amount of false negative interactions (1-4).

Mirroring the experimental techniques, computational methods have also been developed to identify new PPIs. These computational methods can be divided into three main approaches, depending on the contextual properties they exploit: structural, genomic or biological (5). Structural context methods such as InterPreTS (6), PIPE (7) or Struct2Net (8) extrapolate structural information of a protein directly from its sequence and predict or score PPIs based on the molecular composition and structural conformation of the partners. On the contrary, genomic context methods like STRING

(9) or Predictome (10) provide predictions of putative *in vivo* PPIs based on gene fusion, gene co-localization and phylogenetic profiles. Finally, biological context methods (e.g. GeneCensus (11)) integrate several experimental high-throughput datasets and use bayesian networks to produce more reliable interactions. PPI evidences and predictions provided by these methods are compiled in BIANA (12). To properly assess the success of any PPI prediction method, a reference set of non-interacting protein pairs (NIPs) is required (11,13-15). The Negatome database (16) is a set of known non-interacting proteins (NIP). Specifically, it gathers pairs of proteins that are unlikely to engage in physical direct interactions through manually curated literature and crystallographic data.

Here we present the iLoops web-server, a web implementation of the iLoops structural context method (17) that exploits ArchDB classification of loops (18) and SCOP domains (19) to predict PPI. Briefly, the method assigned structural features (either loops or domains) to a pair of protein sequences. It described each protein pair with two sets of *protein signatures* (combinations of up to three structural features, loops or domains) defined as *interaction signatures*. The number of favouring and disfavouring interaction signatures, the likelihood of the best signatures favouring or disfavouring the interaction and the ratios between them were used to train a random forest approach (ref) and generate a predictor model. Finally, in order to apply the method on sets with a large number of non-interactions, the predictor model was trained under different relative-costs that penalized the errors of false positive predictions. .The server uses sequence similarity to assign structural features (either loops or domains) to each pair of query sequences of the input. Then, the server obtains the set of *interaction signatures* to describe the pair of proteins been tested. The server classifies the interaction features as favouring or disfavouring the interaction and it applies the previously trained random forest model using the relative-cost selected by the user.

METHODS AND IMPLEMENTATION

Assignment of protein signatures. Structural features are annotated over the sequences by means of sequence similarity using BLAST (20). Structural features are assigned to a query protein when the percentage of identities (loops) or similarities (domains) of the sequence alignment is above the twilight zone (21) and the coverage of the structural feature is high enough (100% for loops, 75% for domains). Protein signatures of each query sequence are built with all possible combinations of up to three structural features of the same type (see Figure 1.a).

Evaluation of interaction signatures. Interaction signatures between two query proteins are obtained with the combinations of protein signatures from both. All possible interaction signatures assigned to the pair are examined in previously tabulated scoring matrices. These matrices contain the probabilities of observing each signature in PPIs (favouring matrix, M^+) or NIPs (disfavouring matrix, M^-) (17). Then, interaction signatures are denoted as favouring if scored in M^+ or disfavouring if scored in M^- (see Figure 1b).

Predicting a PPI. The server applies a random forest model, previously trained (17) using the WEKA package (22), to test the protein pair (see Figure 1c). Several random forest classifiers were obtained using different relative costs to penalize the ratio of false positive predictions (17). Each relative cost is associated to a certain expectation of success according to the expected unbalance between PPIs and NIPs in the query data, Thus, the user can select in the input the best relative-cost for the set of proteins in the input.

RESULTS: SERVER USAGE

The input for the iLoops web server is a set of FASTA formatted sequences including the title with a protein identifier (PID) and a list of pairs of proteins (two PIDs separated by a double column "::") to test. Data can be provided through a text area or uploading a file. Finally, the user must select the type of structural features to use for the prediction (loop by default) and the relative cost of false positive predictions. Each submission is limited to 25 protein-pairs. The server will provide a job identification code that can be used either to retrieve the predictions through the results-page or as a bookmark.

The predictions can be browsed through the web interface or downloaded in compressed text files. The predictions are provided as a boolean decision for each queried pair from the input list, along with the final score given by the random forest classifier. Details of each prediction can be displayed in a brief summary with the parameters used for the classification, the structural features assigned to each query protein and a list of favouring (positive) and disfavouring (negative) interaction signatures sorted by their p-value.

DISCUSSION AND CONCLUSION

In this work we have presented a server to predict new protein-protein interactions through the identification of structural features. The iLoops server provides a user-friendly interface and a comprehensive results-page. All the results can be traced back to the original databases to provide means to understand how the prediction was carried out. Such traceability allows the user to comprehend the results and devise new experiments that could be relevant for a particular interaction. Additionally, the iLoops server offers to interested researchers the possibility to select the relative cost for false positive discoveries in their required predictions, becoming a unique framework to associate the knowledge on the interacting candidates to fair expectations of prediction success.

ACKNOWLEDGEMENT

The authors want to thank Javier Garcia-Garcia for his useful discussion and help with the search of known Y2H confirmed PPIs using BIANA.

FUNDING

This work was supported by the Spanish Ministry of Science and Innovation (MICINN) [FEDER BIO2008-0205, FEDER BIO2011-22568]. MAML wants to acknowledge the FI-DGR 2012 fellowship from "Generalitat de Catalunya". Funding for open access charge: Spanish Ministry of Science and Innovation.

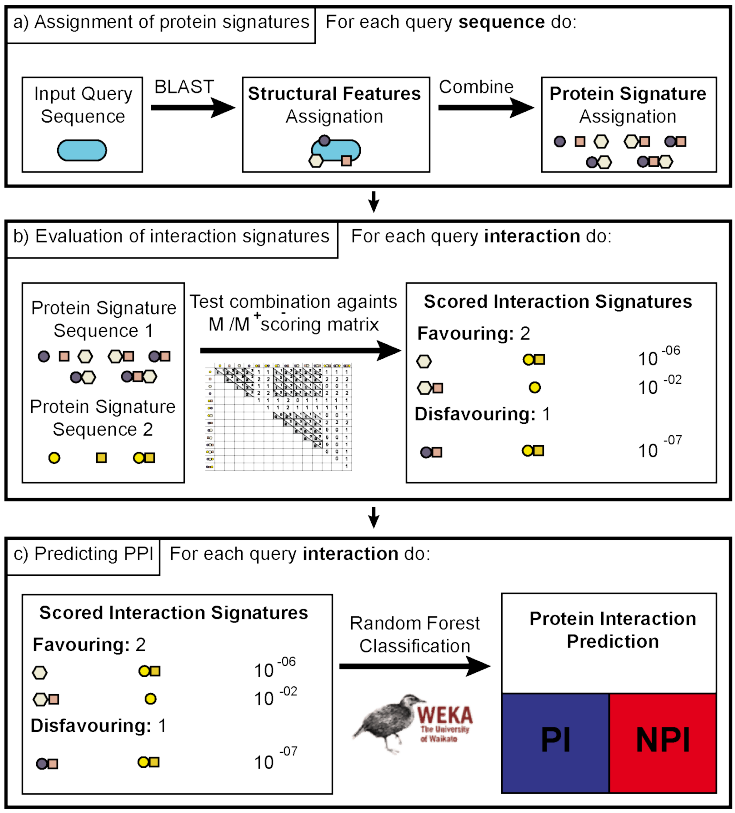
REFERENCES

1. Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104-110.
2. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.S. *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*, **6**, 91-97.
3. Rajagopala, S.V., Hughes, K.T. and Uetz, P. (2009) Benchmarking yeast two-hybrid systems using the interactions of bacterial motility proteins. *Proteomics*, **9**, 5296-5302.
4. Stellberger, T., Hauser, R., Baiker, A., Pothineni, V.R., Haas, J. and Uetz, P. (2010) Improving the yeast two-hybrid system with permutated fusions proteins: the Varicella Zoster Virus interactome. *Proteome Sci*, **8**, 8.
5. Skrabanek, L., Saini, H.K., Bader, G.D. and Enright, A.J. (2008) Computational prediction of protein-protein interactions. *Mol Biotechnol*, **38**, 1-17.
6. Aloy, P. and Russell, R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161-162.
7. Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., Greenblatt, J., Jessulat, M., Krogan, N. *et al.* (2006) PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, **7**, 365.
8. Singh, R., Park, D., Xu, J., Hosur, R. and Berger, B. (2010) Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic Acids Res*, **38**, W508-515.
9. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, **39**, D561-568.
10. Mellor, J.C., Yanai, I., Clodfelter, K.H., Mintseris, J. and DeLisi, C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*, **30**, 306-309.
11. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449-453.
12. Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J. and Oliva, B. (2010) Biana: a software framework for compiling biological interactions and analyzing networks. *BMC bioinformatics*, **11**, 56.
13. Ben-Hur, A. and Noble, W.S. (2006) Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7 Suppl 1**, S2.

14. Trabuco, L.G., Betts, M.J. and Russell, R.B. (2012) Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods*.
15. Yu, J., Guo, M., Needham, C.J., Huang, Y., Cai, L. and Westhead, D.R. (2010) Simple sequence-based kernels do not predict protein-protein interactions. *Bioinformatics*, **26**, 2610-2614.
16. Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D. *et al.* (2010) The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*, **38**, D540-544.
17. Planas-Iglesias, J., Bonet, J., Garcia-Garcia, J., Marín-López, M.A., Feliu, E. and Oliva, B. (2012) Understanding protein-protein interactions using local structural features *Submitted to Journal of molecular biology*.
18. Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F.X., Sternberg, M.J. and Oliva, B. (2004) ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic acids research*, **32**, D185-188.
19. Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*, **36**, D419-425.
20. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.
21. Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*, **12**, 85-94.
22. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, **11**.

Figures

Figure 1: Pipeline of the iLoops server procedure



The pipeline of the iLoops server can be summarized in three steps applied to each pair of proteins of the input list:

a) The assignment of structural features through BLAST similarity search. Protein signatures are defined as groups of up to three structural features.

b) Scores obtained from M^+ and M^- matrices extracted from sets of PPIs and NIPs are assigned to all interaction signatures (pairs of protein signatures) of a protein-pair.

c) A random forest classifier evaluates a set of parameters that describe the interaction signatures, and provides a final prediction for the queried protein pair.

3.3 Extending signalling pathways: application to apoptosis pathways

Planas-Iglesias, J., Guney, E., Garcia-Garcia, J., Robertson, K. A., Raza, S., Freeman, T. C., Ghazal, P. & Oliva, B. (2012). Extending signaling pathways with protein-interaction networks. Application to apoptosis. *OMICS* **16**, 245-56.

Planas-Iglesias J, Guney E, Garcia-Garcia J, Robertson KA, Raza S, Freeman TC, et al. Extending signaling pathways with protein-interaction networks. Application to apoptosis. Supplementary data. OMICS. 2012 May;16(5):245-256.

4 Discussion

4.1 Overview

In this thesis I have adressed the objectives of developing two new methods to predict protein-protein interactions and new members of signalling pathways. Despite the apparent disparity of the objectives, several commonalities among them are worthwhile to be noted.

In first place, the two developed algorithms strongly rely on previous knowledge of protein-protein interactions. Hence, the importance of counting with reference protein interaction networks as complete and accurate as possible is emphasized (see sections 1.1.6 and 1.3.5). To this extent, it is imperative the use of tools for integrating low- and high-throughput experimental data into meaningful biological networks. In this thesis, PIANA (272) and BIANA (92) frameworks for the integration and analysis of biological networks were entrusted such task.

In second term, both methods take advantage of the integration of contextual information to attain accurate predictions. On one hand, results from section 3.2 demonstrate the importance of the structural knowledge of proteins (in the form of local structural features or loops (168)) to improve the performance of current methods for PPI prediction. On the other hand, it is shown in section 3.3 that besides real PPIs, protein functional association information such as that enclosed in STRING database (68) (see table 1.4) is crucial to successfully transfer functional annotation from known apoptosis-related proteins to meaningful new candidates.

Third, the results of both research pieces stress out the necessity of appropriate negative PPI models for PPI prediction. Regarding the prediction of PPIs based on local structural features, the knowledge of real non-interacting pairs is crucial to extract characteristic loop signatures that denote negative interactions (i.e. protein pairs that do not interact). Such information could not be obtained from simulated negative models. In the case of extending the apoptosis pathways, the different methods used transferred an *apoptosis score* from known apoptosis proteins to their interacting partners, either in the real PPI network or in other random ones used as negative reference. Then, an unanotated protein is considered as apoptosis candidate if there exist significant differences between the score obtained in the real network and that obtained in the

random ones. The results presented in section 3.3 show that all methods yielded such a large number of candidates that the obtained predictions are without biological sense. These results suggest that the simulated negative models do not enclose enough information to differentiate between biologically meaningful candidates and the rest of them.

Within the following sections, the main results from the research on PPI prediction based on local structural features (sections 4.2, 4.3, and 4.4), and from the study on the completion of signalling pathways (section 4.6) are further discussed. Besides, the stated necessity for good negative models in network-based protein interaction predictions is argued in the context of the results obtained from both research objectives (section 4.5). Finally, future directions for the research presented in this thesis are discussed in section 4.7.

4.2 Relevance of small local structural features upon the establishment of protein binding

The iLoops algorithm presented in section 3.2 explores the capability of protein loops (as classified in (168)) to explain the mechanism underlying the formation of protein-protein interactions. The algorithm scores the likelihood of a protein-protein interaction by considering the different groups of loops within a protein a footprint of its interacting potential (protein signature). Then, it obtains characteristic interaction signatures as combinations of pairs of protein signatures, where each member of the interaction signature represents one of the proteins in the evaluated protein pair. The score provided by the method corresponds to the probability of observing the interaction signature at least as often as the number of occurrences in a reference set. Depending on whether the reference set is formed by PPIs or experimentally determined non-interacting pairs (NIPs) (271) the signature and its score can be regarded as positive or negative. Although a particular signature can be observed both in PPIs and NIPs, it is observed in the different sets with different frequencies, thus obtaining different positive and negative scores.

The results obtained show that, PPIs and NIPs are characterised by different types of signatures. Both positive and negative signatures are to some extent able to discern between reference sets of PPIs and NIPs. Interestingly, when the ratio between the scores of positive and negative signatures is considered, the differentiation between the reference sets is maximal. These results corroborate recent findings about the crucial role of loops in the formation of PPIs (70,167,273), but also in preventing the formation of protein interaction complexes. Furthermore, the fact that negative interaction signatures are able to discern between PPIs and NIPs shows that relevant information was enclosed in NIPs.

Scoring the likelihood of an interaction based on the scores of positive and negative interaction signatures considered only one interface produced by a putative collision. However, many of such interfaces could potentially occur upon the encounter of two proteins. Hence, the analysis on whether the total number of signatures could be a good criterion of classification was done, achieving a considerably better performance than the interaction signatures scores. These results may be interpreted in the light of a recent study made by Wass et al. (166). In their work, they studied the capability of docking algorithms (which are used to identify the best interface between a pair of proteins) to predict interacting partners regardless of their success in pinpointing the interacting region. Their study showed that although docking algorithms could fail to identify the native complex (and thus, the interface), the distribution of docking scores discerned between interacting and non-interacting pairs. From their results, the authors suggested that protein surface morphology contained sufficient information to identify a bona fide interactor (166). This concept implied that several regions of the protein were important for the molecular association between two proteins. Indeed, this described a model of the funnel-like intermolecular energy landscape in PPIs (274,275).

4.3 The funnel-like intermolecular energy landscape framework

It is unclear how two interacting proteins can find each other within a large population of proteins and quickly form a binary complex. If the molecular association was the

result of a quasi-infinite series of elastic collisions with a unique successful outcome (i.e. the final pose of the binary complex), the formation of PPIs would require an unaffordable time-scale as in Levinthal's paradox (276,277). The solution of the problem is to assume that in the collision of two proteins they recognise if they have to interact or not, forming an intermediate complex that may (or may not) have the best docking interface. For a non-interacting pair, both proteins would be immediately released to interact with others, while if they had to interact they would stay together (or near each other) until finding the correct conformation. This model implies certain "stickiness" between the interacting proteins, which would allow the formation of the intermediate complexes.

In this context, the obtained results suggest a similar explanation for the formation of interacting protein-pairs. Being the number of interaction signatures a good classifier of PPIs and NIPs, it can be inferred that not only one interacting region is important to decide whether a pair of proteins could interact. Hence, several protein regions could participate in the interaction process, allowing the formation of intermediates of the binary complex. In the framework of the funnel-like intermolecular energy landscape theory, proteins would explore their energetic landscape during the protein-protein collisions. From the work herein presented, it can be proposed that this landscape is constrained by the composition of local structural features of both proteins (protein signatures). According to this model, positive and negative interaction signatures could represent energetic valleys and peaks respectively. Thus, the pairing of protein signatures would encode the possibilities to accept or not the interaction, allowing to use such signatures to predict the potential interaction between two proteins.

4.4 Use of local structural features for protein interaction prediction

The aim of predicting PPIs is to help the researcher who wants to test interactions from a random selection of protein pairs of a proteome. Current estimates of the interactome size in human and other model organisms (65,266,268) indicate that the number of non-interacting pairs largely exceeds the number of existing PPIs. This unbalance has a vast

impact on the performance of predictive methods (278,279), but can be mitigated by the expertise of the researcher who asks for the predictions. For instance, protein pairs that are compartmentalised apart in the cell (95% of all protein pairs in the human proteome) would rarely make a suitable set of candidates for protein-protein interaction. Regarding this issue, an independent validation of the iLoops method showed that the decrease of performance associated to an increased unbalance between PPIs and NIPs could be to some extent compensated by penalising the errors produced by false positive predictions (the *relative cost* of the predictions). Compared to other available methods for the prediction of protein interactions (264,265,280-282), iLoops offers to interested researchers the possibility to fix the relative cost of the required predictions, becoming a unique framework to associate the knowledge on the interacting candidates to fair expectations of prediction success. Furthermore, the availability of an easy-to-use web server implementation of iLoops (http://sbi.imim.es/iLoops.php) should permit the scientific community an extensive use of the method.

4.5 Negative protein interaction models: random networks and experimental negative data

One of the key features in iLoops algorithm is the use of the Negatome database (271) as a negative model for PPIs. However, regarding PPI prediction the requirements of a good negative model are still under discussion (283). On one hand, it is still nowadays difficult to identify non-interacting protein pairs due to the lack of sensitivity of high-throughput experimental methods for PPI detection (65,234). On the other hand, it has been demonstrated that random negative models may introduce several biases, depending upon their required topology, the unbalance with respect to the number of PPIs assumed, or the compartmentalisation of the protein pairs the negative model encloses (263,278). Despite the experimental bias of the Negatome database (284), the obtained results show that non-interacting pairs in the Negatome database contain relevant structural information for discerning between PPIs and NIPs. How can this apparent contradiction be solved? It has been shown that binding residues of a PPI are subject to co-evolution constraints (202) imposed by the needs of the interacting

partners to “tell” each other they have to interact. Several methods have taken advantage of such constraints to predict PPIs (196,203). If non-interacting pairs had analogous requirements (i.e. to expose signals that prevented their interaction), similar constraints would apply to NIPs and could be exploited for their prediction. According to the funnel-like intermolecular energy landscape theory (see section 4.3), this should be the case.

Interestingly, results from the study on extending apoptosis signalling pathways also hinted the limitations of simulated negative interaction data. Such random networks lead different methods for functional annotation transfer to yield an outsized and biologically meaningless number of predictions. In this case, the results can be understood at the light of a topological explanation. Random networks can be designed to preserve centrality and modularity properties similar to the ones a real interactome has (284). However, functional annotation transfer methods rely not only in the overall topology of the network, but also in the fact that important proteins for the studied system or phenotype are central (i.e. hubs) in the real network (44,45,284). Random networks cannot grant both the same centrality degree of such proteins and being different enough to actual data at the same time. Since differentiation from real interactions takes precedence in the construction of negative models, the centrality of key proteins is to some extent diminished. Thus, random networks artificially enlarge differences between the negative and the real models. Due to the lack of negative data, simulated negative models have been the only feasible alternative to exploit functional annotation transfer methods during long time. However, recently developed methods for extracting negative data from high-throughput experiments (269) may help to surmount this problem.

All together, it can be deduced from both research pieces that the disposal of an appropriate negative model is crucial to obtain accurate PPI predictions.

4.6 Functional annotation transfer: lessons from a controlled retrospective experiment

The negative impact of using randomly generated networks as a negative model in annotation transfer methods can be measured with a retrospective experiment. Section 3.3 describes how different scoring methods were used to transfer annotation from 53 well-studied members of the human apoptosis pathways (as known by 2005) to their protein interactors. The selected scoring methods base their predictions in different proximity measures, either direct neighbourhood (285) or shortest paths (50,52) (see section 1.1.4). All scoring methods produced significant predictions (compared to a random negative model), but its number was too large to be useful (see previous section). To approximate the overestimation of score produced by the use of random networks as negative models, the results of the different methods analysed were compared to a validation set conformed by the proteins newly related to the apoptosis pathways in the period 2005-2010. Based on the different methods reporting a given prediction and the overlap between predictions and the validation set, a method was developed to score the reliability of predictions not present in the validation set, which potentially could be relevant in the apoptosis pathways.

To better understand the applicability of functional annotation methods to signalling pathways, four different reference networks were used to extend the apoptosis annotation. These networks enclosed incremental amounts of information:

i) experimentally determined PPIs obtained from PIANA (272);

ii) functional associations from STRING database (68);

iii and iv) interology predictions (190) over the two preceding networks.

Comparing the results of different functional annotation methods applied to the described reference networks, the relevance of functional associations for pinpointing typical elements from signalling pathways such as phosphorilation events was observed. These results are consistent with previously reported ones (81), and are in consonance

with observations from Valente *et al.* who reported that high-throughput methods for PPI detection contribute little new knowledge about phosphorilation events and interactions (79). Furthermore, these findings are in direct agreement with recent experiments by Breitkreutz *et al.*, who showed that specific experimental approaches are required to pinpoint the transient nature of most PPIs in signalling pathways (80).

Finally, the previously described approach for scoring the reliability of predictions was applied to the results obtained when using the reference networks. From the 53 well-studied members of the human apoptosis pathways, the total number of predictions yielded by the scoring methods using the interaction data as known by 2005 was in the scale of thousands. To measure the impact of using random networks as negative models in the prediction scores, these predictions were compared to a validation set conformed by the proteins newly related to the apoptosis pathways in the period 2005-2010. From the initial set of predictions, only 273 matched reliability criteria to be selected as relevant candidates for the apoptosis pathways. These results indicate that methods for functional annotation transfer overestimate their scores due to the use of random negative data, leading to oversized and biologically meaningless predictions. Although most of the selected candidates were totally unannotated, a functional trend enrichment analysis (286) revealed that the functional annotation of the few remaining candidates was compatible with their potential role as apoptosis-related proteins.

4.7 Further directions

Different perspectives arise for the two methods presented in this thesis. From the development point of view, the study on the extension of signalling pathways is self-enclosed. The method can be applied to nay other cell process or phenotype provided the availability of time-labelled data: PPI networks obtained from earlier data and validation sets representing the present knowledge of the studied cell process or phenotype. However, there is a great scope for further experimental analyses on the proposed candidates.

Regarding iLoops method for protein interactions prediction, there is larger space for improvement. First, the method strongly relies in the structural classification of loops

(168), which is slightly out of date. The current classification is under revision and a prompt update is already scheduled. Second, iLoops users could take benefit from the possibility of applying a wider variety of structural features, including Pfam (287) and CATH (147) definitions of domains. Such improvement is also being implemented. Third, the evidences found hinting the specific and important role of loops in the protein binding process in the framework of the funnel-like intermolecular energy landscape theory open space for new research perspectives. Can loop interaction signatures (their number or their scores) recapitulate the funnel-like energy landscape on the surface of the interacting proteins? If so, can loop interaction signatures be used to identify binding interfaces? Two different experiments have been designed to answer these questions. Using either PRISM defined interfaces (288) or docking poses from binding and non-binding partners (166) as reference for the definition of the interacting region, the role of loops and loop interaction signatures in the molecular association of proteins can be further investigated.

5 Conclusions

The main contributions or this thesis can be summarized as follows:

i) The relevance of local structural features such as protein loops on the protein binding process has been statistically assessed. For a given combination of loop groups in a protein pair (interaction signature) the exact probability of observing such signature in reference sets of interacting and non-interacting pairs was computed.

ii) Protein-protein interactions and non-interacting pairs are characterised by different types of interaction signatures. The two different scores obtained from PPIs and NIPs reference sets can discern between interacting and non-interacting pairs. Considering both scores altogether, the differentiation between PPIs and NIPs is maximized.

iii) Each interaction signature considers one interface produced by a putative protein-protein collision. Depending upon the origin of the signature (PPIs or NIPs reference sets), it can be considered that the interface favours or hampers the binding process. The number of such interfaces, defined by the number of interaction signatures, is also a good PPI predictor.

iv) These observations strongly support the funnel-like intermolecular energy landscape theory, which imply that binding proteins explore each other surface, early recognising their interacting potential prior to reaching the final docking state.

v) Based on the stated observations, iLoops, a new method for predicting protein-protein interactions, has been developed. Considering the natural unbalance between PPIs and NIPs, and applying different false discovery costs to the predictions, the method provides a unique framework to researchers for associating their knowledge of on the interacting candidates to fair expectations of prediction success.

vi) Methods for functional annotation transfer overestimate their scores due to the use of random negative data, leading to oversized and biologically meaningless predictions. These results are in agreement with recent findings

by Trabuco *et al*. (269). In the framework of a retrospective controlled experiment, a new method for quantifying such overestimate has been developed.

vii) The method was applied to the prediction of new members of the human apoptosis pathways, allowing the selection of 273 reliable candidates over thousands of predictions yielded by different functional annotation transfer methods.

viii) Among different types of methods for functional annotation transfer, direct neighbourhood based methods can be applied to small PPI networks; however an accurate annotation transfer in larger networks requires other approaches such as those based in shortest paths definitions for node proximity in the network.

ix) Further insight on the nature of signalling networks was gained by appliying functional annotation transfer methods to different reference PPI networks, which included experimentally PPIs, functional associations, and interology predictions. The obtained results confirmed that classical high-throughput techniques for PPI detection contribute little new knowledge about the phosphorilation events and interactions characteristic of signalling pathways.

6 Appendix

6.1 Overview

In this appendix, contributions I made in other pieces of research unrelated ot this thesis are presented. Briefly, I participated in two additional publications:

i) Biana: A software framework for compiling biological interactions and analyzing networks

ii) Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details

In the first paper, I collaborated in the design of the protocol that BIANA uses for unifying biological entity identifiers, a central tool for integrating biological data from different sources. The protocol allows the user to indicate a set of identifiers to be used for the unification of different records from external databases. In this paper, I also contributed the presented example on the reconstruction of metabolic networks. The reconstruction was obtained by chaining reactions between enzymes A and B whenever there is at least one chemical compound in the intersection, this is acting at the same time as product of enzyme A and substrate of enzyme B. Then, chained reactions were scored according to the plausibility of observing chemical compounds in the intersection, taking into account their own frequency and the frequency of other products of enzyme A and other substrates of enzyme B that did not contribute to the intersection.

The second publication is an extensive review on protein-protein interactions, the different levels of detail at which PPIs can be studied, methods to detect and predict them, repositories for PPI data, and functional relationships between the different detail levels in protein-protein interactions. My contribution was centered in designing the structure of the paper contents, the literature review, and the critical assessment of the final manuscript.

Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J, Oliva B. Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics. 2010 Jan 27;11:56-2105-11-56.

Garcia-Garcia, J., Bonet, J., Guney, E., Fornes, O., Planas, J. and Oliva , B. Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details. Mol Inf. 201; 31: 342 -362.

Bibliography.

1. Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561-563.
2. Crick, F.H. (1958) On protein synthesis. *Symp Soc Exp Biol*, **12**, 138-163.
3. Li, G.W. and Xie, X.S. (2011) Central dogma at the single-molecule level in living cells. *Nature*, **475**, 308-315.
4. Spitz, F. and Furlong, E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*.
5. Eaton, W.A., Munoz, V., Hagen, S.J., Jas, G.S., Lapidus, L.J., Henry, E.R. and Hofrichter, J. (2000) Fast kinetics and mechanisms in protein folding. *Annu Rev Biophys Biomol Struct*, **29**, 327-359.
6. Rose, G.D., Fleming, P.J., Banavar, J.R. and Maritan, A. (2006) A backbone-based theory of protein folding. *Proc Natl Acad Sci U S A*, **103**, 16623-16633.
7. Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T. *et al.* (2007) Determining the architectures of macromolecular assemblies. *Nature*, **450**, 683-694.
8. Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J. and Marti-Renom, M.A. (2010) The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*, **18**, 107-114.
9. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631-636.
10. Dohlman, H.G. and Thorner, J.W. (2001) Regulation of G protein-initiated signal transduction in yeast: paradigms and principles. *Annu Rev Biochem*, **70**, 703-754.
11. Stark, G.R. and Darnell, J.E., Jr. (2012) The JAK-STAT pathway at twenty. *Immunity*, **36**, 503-514.
12. Ljungdahl, P.O. and Daignan-Fornier, B. (2012) Regulation of amino acid, nucleotide, and phosphate metabolism in Saccharomyces cerevisiae. *Genetics*, **190**, 885-929.
13. Cooper, T. (1982) In Strathern, J. N., Jones, E. W. and Broach, J. R. (eds.), *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 39-100.
14. Jones, E.W. and Fink, G.R. (1982) In Strathern, J. N., Jones, E. W. and Broach, J. R. (eds.), *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expresion*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 181-299.
15. Ravasz, E. (2009) Detecting hierarchical modularity in biological networks. *Methods Mol Biol*, **541**, 145-160.
16. Committee, B.D. (2000).
17. Gross, J. and Yellen, J. (1999) *Graph Theory and Its Applications*. CRC Press, Boca Raton.

18. Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**, 101-113.
19. Edrös, P. and Renyi, A. (1960) On the evolution of random graphs. . *Publ. Math. Inst. Hung. Acad. Sci.* , **5**, 17-61.
20. Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509-512.
21. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651-654.
22. Wagner, A. and Fell, D.A. (2001) The small world inside large metabolic networks. *Proc Biol Sci*, **268**, 1803-1810.
23. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727-1736.
24. Yook, S.H., Oltvai, Z.N. and Barabasi, A.L. (2004) Functional and topological characterization of protein interaction networks. *Proteomics*, **4**, 928-942.
25. Hintze, A. and Adami, C. (2008) Evolution of complex modular biological networks. *PLoS Comput Biol*, **4**, e23.
26. Lima-Mendez, G. and van Helden, J. (2009) The powerful law of the power law and other myths in network biology. *Mol Biosyst*, **5**, 1482-1493.
27. Arda, H.E. and Walhout, A.J. (2010) Gene-centered regulatory networks. *Brief Funct Genomics*, **9**, 4-12.
28. Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A.M., Grove, C.A., Martinez, N.J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J.S., Hope, I.A. *et al.* (2006) A gene-centered C. elegans protein-DNA interaction network. *Cell*, **125**, 1193-1205.
29. Featherstone, D.E. and Broadie, K. (2002) Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioessays*, **24**, 267-274.
30. Ma'ayan, A., Jenkins, S.L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N.J., Weng, G., Ram, P.T., Rice, J.J. *et al.* (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science*, **309**, 1078-1083.
31. Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47-52.
32. Kashtan, N. and Alon, U. (2005) Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci U S A*, **102**, 13773-13778.
33. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824-827.
34. Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, **31**, 64-68.
35. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551-1555.
36. Oliver, S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601-603.
37. Quackenbush, J. (2001) Computational analysis of microarray data. *Nat Rev Genet*, **2**, 418-427.

38. Quackenbush, J. (2003) Genomics. Microarrays--guilt by association. *Science*, **302**, 240-241.
39. Wu, L.F., Hughes, T.R., Davierwala, A.P., Robinson, M.D., Stoughton, R. and Altschuler, S.J. (2002) Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nat Genet*, **31**, 255-265.
40. Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat Biotechnol*, **18**, 1257-1261.
41. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623-627.
42. Ulitsky, I., Maron-Katz, A., Shavit, S., Sagir, D., Linhart, C., Elkon, R., Tanay, A., Sharan, R., Shiloh, Y. and Shamir, R. (2010) Expander: from expression microarrays to networks and functions. *Nat Protoc*, **5**, 303-322.
43. Sharan, R., Ulitsky, I. and Shamir, R. (2007) Network-based prediction of protein function. *Mol Syst Biol*, **3**, 88.
44. Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41-42.
45. Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, **100**, 12123-12128.
46. Belkhadir, Y. and Chory, J. (2006) Brassinosteroid signaling: a paradigm for steroid hormone signaling from the cell surface. *Science*, **314**, 1410-1411.
47. Ehebauer, M., Hayward, P. and Arias, A.M. (2006) Notch, a universal arbiter of cell fate decisions. *Science*, **314**, 1414-1415.
48. Slessareva, J.E. and Dohlman, H.G. (2006) G protein signaling in yeast: new components, new connections, new compartments. *Science*, **314**, 1412-1413.
49. Kim, P.M., Lu, L.J., Xia, Y. and Gerstein, M.B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, **314**, 1938-1941.
50. Guney, E. and Oliva, B. (2012) Exploiting Protein-Protein Interaction Networks for Genome-wide Disease-Gene Prioritization. *PLoS One*, **In press.**
51. Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tumer, Z., Pociot, F., Tommerup, N. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, **25**, 309-316.
52. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. and Singh, M. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21 Suppl 1**, i302-310.
53. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. and Sharan, R. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*, **6**, e1000641.
54. Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Res*, **18**, 644-652.
55. Vidal, M., Cusick, M.E. and Barabasi, A.L. (2011) Interactome networks and human disease. *Cell*, **144**, 986-998.

56. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. and Marcotte, E.M. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*, **21**, 1109-1121.
57. Navlakha, S. and Kingsford, C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**, 1057-1063.
58. Bar-Joseph, Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493-2503.
59. Werhli, A.V., Grzegorczyk, M. and Husmeier, D. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, **22**, 2523-2531.
60. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
61. Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009) Partial Correlation Estimation by Joint Sparse Regression Models. *J Am Stat Assoc*, **104**, 735-746.
62. Schafer, J. and Strimmer, K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754-764.
63. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7 Suppl 1**, S7.
64. Allen, J.D., Xie, Y., Chen, M., Girard, L. and Xiao, G. (2012) Comparing statistical methods for constructing large scale gene networks. *PLoS One*, **7**, e29348.
65. Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat Methods*, **6**, 83-90.
66. Bader, J.S., Chaudhuri, A., Rothberg, J.M. and Chant, J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, **22**, 78-85.
67. Kamburov, A., Stelzl, U. and Herwig, R. (2012) IntScore: a web tool for confidence scoring of biological interactions. *Nucleic Acids Res*, **40**, W140-146.
68. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, **39**, D561-568.
69. Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, **311**, 681-692.
70. Sprinzak, E., Altuvia, Y. and Margalit, H. (2006) Characterization and prediction of protein-protein interactions within and between complexes. *Proc Natl Acad Sci U S A*, **103**, 14718-14723.

71. Stein, A., Ceol, A. and Aloy, P. (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res*, **39**, D718-723.
72. Lehner, B. and Fraser, A.G. (2004) A first-draft human protein-interaction map. *Genome Biol*, **5**, R63.
73. Aragues, R., Sali, A., Bonet, J., Marti-Renom, M.A. and Oliva, B. (2007) Characterization of protein hubs by inferring interacting motifs from protein interactions. *PLoS Comput Biol*, **3**, 1761-1771.
74. Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. and Chinnaiyan, A.M. (2005) Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, **23**, 951-959.
75. Albrecht, M., Huthmacher, C., Tosatto, S.C. and Lengauer, T. (2005) Decomposing protein networks into domain-domain interactions. *Bioinformatics*, **21 Suppl 2**, ii220-221.
76. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U. (2004) Superfamilies of evolved and designed networks. *Science*, **303**, 1538-1542.
77. Gandhi, T.K., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B. *et al.* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, **38**, 285-293.
78. Barrios-Rodiles, M., Brown, K.R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R.S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I.W. *et al.* (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. *Science*, **307**, 1621-1625.
79. Valente, A.X., Roberts, S.B., Buck, G.A. and Gao, Y. (2009) Functional organization of the yeast proteome by a yeast interactome map. *Proc Natl Acad Sci U S A*, **106**, 1490-1495.
80. Breitkreutz, A., Choi, H., Sharom, J.R., Boucher, L., Neduva, V., Larsen, B., Lin, Z.Y., Breitkreutz, B.J., Stark, C., Liu, G. *et al.* (2010) A global protein kinase and phosphatase interaction network in yeast. *Science*, **328**, 1043-1046.
81. Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A., Jorgensen, C., Miron, I.M., Diella, F., Colwill, K., Taylor, L., Elder, K. *et al.* (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415-1426.
82. Navlakha, S., Gitter, A. and Bar-Joseph, Z. (2012) A Network-based Approach for Predicting Missing Pathway Interactions. *PLoS Comput Biol*, **8**, e1002640.
83. Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*, **39**, D691-697.
84. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, **40**, D109-114.
85. Oberhardt, M.A., Palsson, B.O. and Papin, J.A. (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, **5**, 320.

86. Pearson, H. (2001) Biology's name game. *Nature*, **411**, 631-632.
87. Orchard, S. and Hermjakob, H. (2008) The HUPO proteomics standards initiative--easing communication and minimizing data loss in a changing world. *Brief Bioinform*, **9**, 166-173.
88. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C. *et al.* (2004) The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol*, **22**, 177-183.
89. Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat Biotechnol*, **28**, 935-942.
90. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524-531.
91. Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T.P. and Hautaniemi, S. (2009) Integrated network analysis platform for protein-protein interactions. *Nat Methods*, **6**, 75-77.
92. Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J. and Oliva, B. (2010) Biana: a software framework for compiling biological interactions and analyzing networks. *BMC bioinformatics*, **11**, 56.
93. Prieto, C. and De Las Rivas, J. (2006) APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res*, **34**, W298-302.
94. Razick, S., Magklaras, G. and Donaldson, I.M. (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, **9**, 405.
95. Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S., Ceol, A., Chautard, E., Dana, J.M., De Las Rivas, J., Dumousseau, M., Galeota, E. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods*, **8**, 528-529.
96. EBI.
97. Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P. and Philippi, S. (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, **22**, 1383-1390.
98. Garcia-Garcia, J., Bonet, J., Guney, E., Fornes, O., Planas, J. and Oliva, B. (2012) Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details. *Molecular Informatics*, **31**, 342-362.
99. Chitwood, D.H. and Timmermans, M.C. (2010) Small RNAs are on the move. *Nature*, **467**, 415-419.
100. Djuranovic, S., Nahvi, A. and Green, R. (2011) A parsimonious model for gene regulation by miRNAs. *Science*, **331**, 550-553.
101. Dawson, J.H. (1988) Probing structure-function relations in heme-containing oxygenases and peroxidases. *Science*, **240**, 433-439.
102. Stuhmer, W., Conti, F., Suzuki, H., Wang, X.D., Noda, M., Yahagi, N., Kubo, H. and Numa, S. (1989) Structural parts involved in activation and inactivation of the sodium channel. *Nature*, **339**, 597-603.

103. Sturm, R.A. and Herr, W. (1988) The POU domain is a bipartite DNA-binding structure. *Nature*, **336**, 601-604.
104. Watson, J.D., Laskowski, R.A. and Thornton, J.M. (2005) Predicting protein function from sequence and structural data. *Curr Opin Struct Biol*, **15**, 275-284.
105. Anfinsen, C.B., Haber, E., Sela, M. and White, F.H., Jr. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A*, **47**, 1309-1314.
106. Pauling, L. and Corey, R.B. (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A*, **37**, 251-256.
107. Branden, C. and Tooze, J. (1991). Garland, New York.
108. Wetlaufer, D.B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, **70**, 697-701.
109. Bork, P. (1991) Shuffled domains in extracellular proteins. *FEBS Lett*, **286**, 47-54.
110. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Phillips, D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662-666.
111. Chothia, C. and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J*, **5**, 823-826.
112. Lesk, A.M. and Chothia, C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol*, **136**, 225-270.
113. Rueda, M., Chacon, P. and Orozco, M. (2007) Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, **15**, 565-575.
114. Dobbins, S.E., Lesk, V.I. and Sternberg, M.J. (2008) Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci U S A*, **105**, 10390-10395.
115. Meszaros, B., Tompa, P., Simon, I. and Dosztanyi, Z. (2007) Molecular principles of the interactions of disordered proteins. *J Mol Biol*, **372**, 549-561.
116. Tompa, P. (2002) Intrinsically unstructured proteins. *Trends Biochem Sci*, **27**, 527-533.
117. Drenth, J. (2007) *Principles of Protein X-ray Crystallography*. Third ed. Springer Science + Business Media, LLC, New York.
118. Rupp, B. (2010) *Biomolecular Crystallography*. Garland Science, New York.
119. Hunter, M.S. and Fromme, P. (2011) Toward structure determination using membrane-protein nanocrystals and microcrystals. *Methods*, **55**, 387-404.
120. Branden, C. and Tooze, J. (1999) *Introduction to protein structure*. Second ed. Garland New York.
121. Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H.D. and Huber, R. (1997) Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature*, **386**, 463-471.

122. Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**, 905-920.
123. Braun, W., Wider, G., Lee, K.H. and Wuthrich, K. (1983) Conformation of glucagon in a lipid-water interphase by 1H nuclear magnetic resonance. *J Mol Biol*, **169**, 921-948.
124. Fiaux, J., Bertelsen, E.B., Horwich, A.L. and Wuthrich, K. (2002) NMR analysis of a 900K GroEL GroES complex. *Nature*, **418**, 207-211.
125. Frank, J. (2002) Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu Rev Biophys Biomol Struct*, **31**, 303-319.
126. Nickell, S., Kofler, C., Leis, A.P. and Baumeister, W. (2006) A visual approach to proteomics. *Nat Rev Mol Cell Biol*, **7**, 225-230.
127. Schneidman-Duhovny, D., Kim, S.J. and Sali, A. (2012) Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct Biol*, **12**, 17.
128. Judge, P.J. and Watts, A. (2011) Recent contributions from solid-state NMR to the understanding of membrane protein structure and function. *Curr Opin Chem Biol*, **15**, 690-695.
129. Grunewald, K., Desai, P., Winkler, D.C., Heymann, J.B., Belnap, D.M., Baumeister, W. and Steven, A.C. (2003) Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science*, **302**, 1396-1398.
130. Maimon, T., Elad, N., Dahan, I. and Medalia, O. (2012) The human nuclear pore complex as revealed by cryo-electron tomography. *Structure*, **20**, 998-1006.
131. Kurner, J., Frangakis, A.S. and Baumeister, W. (2005) Cryo-electron tomography reveals the cytoskeletal structure of Spiroplasma melliferum. *Science*, **307**, 436-438.
132. Bernado, P. and Svergun, D.I. (2012) Analysis of Intrinsically Disordered Proteins by Small-Angle X-ray Scattering. *Methods Mol Biol*, **896**, 107-122.
133. Krepkiy, D., Mihailescu, M., Freites, J.A., Schow, E.V., Worcester, D.L., Gawrisch, K., Tobias, D.J., White, S.H. and Swartz, K.J. (2009) Structure and hydration of membranes embedded with voltage-sensing domains. *Nature*, **462**, 473-479.
134. Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, **29**, 291-325.
135. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.
136. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.
137. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755-763.
138. Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*, **12**, 85-94.

139. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, **31**, 3497-3500.
140. Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**, 205-217.
141. Fernandez-Fuentes, N., Rai, B.K., Madrid-Aliste, C.J., Fajardo, J.E. and Fiser, A. (2007) Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics*, **23**, 2558-2565.
142. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. and Sali, A. (2007) Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*, **Chapter 2**, Unit 2 9.
143. Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. and Karplus, M. (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins*, **23**, 318-326.
144. Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543-544.
145. Orengo, C.A., Flores, T.P., Taylor, W.R. and Thornton, J.M. (1993) Identification and classification of protein fold families. *Protein Eng*, **6**, 485-500.
146. Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*, **36**, D419-425.
147. Cuff, A.L., Sillitoe, I., Lewis, T., Clegg, A.B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J. and Orengo, C.A. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res*, **39**, D420-426.
148. Bau, D. and Marti-Renom, M.A. (2011) Structure determination of genomic domains by satisfaction of spatial restraints. *Chromosome Res*, **19**, 25-35.
149. Graslund, S., Nordlund, P., Weigelt, J., Hallberg, B.M., Bray, J., Gileadi, O., Knapp, S., Oppermann, U., Arrowsmith, C., Hui, R. *et al.* (2008) Protein production and purification. *Nat Methods*, **5**, 135-146.
150. Bowie, J.U., Luthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-170.
151. Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86-89.
152. Dill, K.A., Ozkan, S.B., Shell, M.S. and Weikl, T.R. (2008) The protein folding problem. *Annu Rev Biophys*, **37**, 289-316.
153. van Gunsteren, W.F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D.P., Glattli, A., Hunenberger, P.H. *et al.* (2006) Biomolecular modeling: Goals, problems, perspectives. *Angew Chem Int Ed Engl*, **45**, 4064-4092.
154. Chiti, F. and Dobson, C.M. (2009) Amyloid formation by globular proteins under native conditions. *Nat Chem Biol*, **5**, 15-22.
155. Wolynes, P.G. (2005) Recent successes of the energy landscape theory of protein folding and function. *Q Rev Biophys*, **38**, 405-410.

156. Simons, K.T., Kooperberg, C., Huang, E. and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, **268**, 209-225.
157. Cozzetto, D. and Tramontano, A. (2008) Advances and pitfalls in protein structure prediction. *Curr Protein Pept Sci*, **9**, 567-577.
158. Bystroff, C., Simons, K.T., Han, K.F. and Baker, D. (1996) Local sequence-structure correlations in proteins. *Curr Opin Biotechnol*, **7**, 417-421.
159. Han, K.F. and Baker, D. (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci U S A*, **93**, 5814-5818.
160. Das, R. and Baker, D. (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem*, **77**, 363-382.
161. Thompson, J. and Baker, D. (2011) Incorporation of evolutionary information into Rosetta comparative modeling. *Proteins*, **79**, 2380-2388.
162. Warner, L.R., Varga, K., Lange, O.F., Baker, S.L., Baker, D., Sousa, M.C. and Pardi, A. (2011) Structure of the BamC two-domain protein obtained by Rosetta with a limited NMR data set. *J Mol Biol*, **411**, 83-95.
163. Joseph, A.P., Valadie, H., Srinivasan, N. and de Brevern, A.G. (2012) Local structural differences in homologous proteins: specificities in different SCOP classes. *PLoS One*, **7**, e38805.
164. Russell, R.B., Sasieni, P.D. and Sternberg, M.J. (1998) Supersites within superfolds. Binding site similarity in the absence of homology. *J Mol Biol*, **282**, 903-918.
165. Fernandez-Recio, J., Totrov, M., Skorodumov, C. and Abagyan, R. (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins*, **58**, 134-143.
166. Wass, M.N., Fuentes, G., Pons, C., Pazos, F. and Valencia, A. (2011) Towards the prediction of protein interaction partners using physical docking. *Molecular systems biology*, **7**, 469.
167. Akiva, E., Itzhaki, Z. and Margalit, H. (2008) Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc Natl Acad Sci U S A*, **105**, 13292-13297.
168. Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F.X., Sternberg, M.J. and Oliva, B. (2004) ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic acids research*, **32**, D185-188.
169. Burke, D.F., Deane, C.M. and Blundell, T.L. (2000) Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics*, **16**, 513-519.
170. Espadaler, J., Querol, E., Aviles, F.X. and Oliva, B. (2006) Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics*, **22**, 2237-2243.
171. Aloy, P. and Russell, R.B. (2004) Ten thousand interactions for the molecular biologist. *Nat Biotechnol*, **22**, 1317-1321.

172. Michnick, S.W. (2001) Exploring protein interactions by interaction-induced folding of proteins from complementary peptide fragments. *Curr Opin Struct Biol*, **11**, 472-477.
173. Ratushny, V. and Golemis, E. (2008) Resolving the network of cell signaling pathways using the evolving yeast two-hybrid system. *Biotechniques*, **44**, 655-662.
174. Lemmens, I., Eyckerman, S., Zabeau, L., Catteeuw, D., Vertenten, E., Verschueren, K., Huylebroeck, D., Vandekerckhove, J. and Tavernier, J. (2003) Heteromeric MAPPIT: a novel strategy to study modification-dependent protein-protein interactions in mammalian cells. *Nucleic Acids Res*, **31**, e75.
175. Russ, W.P. and Engelman, D.M. (1999) TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc Natl Acad Sci U S A*, **96**, 863-868.
176. Stefan, E., Aquin, S., Berger, N., Landry, C.R., Nyfeler, B., Bouvier, M. and Michnick, S.W. (2007) Quantification of dynamic protein complexes using Renilla luciferase fragment complementation applied to protein kinase A activities in vivo. *Proc Natl Acad Sci U S A*, **104**, 16916-16921.
177. Magliery, T.J., Wilson, C.G., Pan, W., Mishler, D., Ghosh, I., Hamilton, A.D. and Regan, L. (2005) Detecting protein-protein interactions with a green fluorescent protein fragment reassembly trap: scope and mechanism. *J Am Chem Soc*, **127**, 146-157.
178. Hu, C.D., Chinenov, Y. and Kerppola, T.K. (2002) Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. *Mol Cell*, **9**, 789-798.
179. Morell, M., Espargaro, A., Aviles, F.X. and Ventura, S. (2008) Study and selection of in vivo protein interactions by coupling bimolecular fluorescence complementation and flow cytometry. *Nat Protoc*, **3**, 22-33.
180. Day, R.N., Periasamy, A. and Schaufele, F. (2001) Fluorescence resonance energy transfer microscopy of localized protein interactions in the living cell nucleus. *Methods*, **25**, 4-18.
181. Xu, Y., Piston, D.W. and Johnson, C.H. (1999) A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins. *Proc Natl Acad Sci U S A*, **96**, 151-156.
182. Soderberg, O., Gullberg, M., Jarvius, M., Ridderstrale, K., Leuchowius, K.J., Jarvius, J., Wester, K., Hydbring, P., Bahram, F., Larsson, L.G. *et al.* (2006) Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nat Methods*, **3**, 995-1000.
183. MacBeath, G. and Schreiber, S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760-1763.
184. Boozer, C., Kim, G., Cong, S., Guan, H. and Londergan, T. (2006) Looking towards label-free biomolecular interaction analysis in a high-throughput format: a review of new surface plasmon resonance technologies. *Curr Opin Biotechnol*, **17**, 400-405.
185. Sprinzak, E., Sattath, S. and Margalit, H. (2003) How reliable are experimental protein-protein interaction data? *J Mol Biol*, **327**, 919-923.

186. Stelzl, U. and Wanker, E.E. (2006) The value of high quality protein-protein interaction networks for systems biology. *Curr Opin Chem Biol*, **10**, 551-558.
187. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751-753.
188. Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, **23**, 324-328.
189. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**, 4285-4288.
190. Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*, **14**, 1107-1118.
191. Wojcik, J. and Schachter, V. (2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **17 Suppl 1**, S296-305.
192. Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A. (1997) Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, **271**, 511-523.
193. Juan, D., Pazos, F. and Valencia, A. (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci U S A*, **105**, 934-939.
194. Moal, I.H. and Bates, P.A. (2012) Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. *PLoS Comput Biol*, **8**, e1002351.
195. Aloy, P. and Russell, R.B. (2006) Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol*, **7**, 188-197.
196. Valdar, W.S. and Thornton, J.M. (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108-124.
197. Hoskins, J., Lovell, S. and Blundell, T.L. (2006) An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci*, **15**, 1017-1029.
198. Jones, S. and Thornton, J.M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol*, **272**, 133-143.
199. Ofran, Y. and Rost, B. (2003) Predicted protein-protein interaction sites from local sequence information. *FEBS Lett*, **544**, 236-239.
200. Ofran, Y. and Rost, B. (2007) ISIS: interaction sites identified from sequence. *Bioinformatics*, **23**, e13-16.
201. Segura, J., Jones, P.F. and Fernandez-Fuentes, N. (2011) Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinformatics*, **12**, 352.
202. Gobel, U., Sander, C., Schneider, R. and Valencia, A. (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309-317.

203. Halperin, I., Wolfson, H. and Nussinov, R. (2006) Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins*, **63**, 832-845.
204. Henschel, A., Winter, C., Kim, W.K. and Schroeder, M. (2007) Using structural motif descriptors for sequence-based binding site prediction. *BMC Bioinformatics*, **8 Suppl 4**, S5.
205. Bogan, A.A. and Thorn, K.S. (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol*, **280**, 1-9.
206. Hu, Z., Ma, B., Wolfson, H. and Nussinov, R. (2000) Conservation of polar residues as hot spots at protein interfaces. *Proteins*, **39**, 331-342.
207. Kundrotas, P.J. and Alexov, E. (2007) PROTCOM: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Res*, **35**, D575-579.
208. Finn, R.D., Marshall, M. and Bateman, A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410-412.
209. Tuncbag, N., Gursoy, A., Nussinov, R. and Keskin, O. (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc*, **6**, 1341-1354.
210. Wodak, S.J. and Janin, J. (1978) Computer analysis of protein-protein interaction. *J Mol Biol*, **124**, 323-342.
211. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A*, **89**, 2195-2199.
212. Gabb, H.A., Jackson, R.M. and Sternberg, M.J. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*, **272**, 106-120.
213. Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R. and Weng, Z. (2007) Integrating statistical pair potentials into protein complex prediction. *Proteins*, **69**, 511-520.
214. Kozakov, D., Brenke, R., Comeau, S.R. and Vajda, S. (2006) PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, **65**, 392-406.
215. Garzon, J.I., Lopez-Blanco, J.R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J. and Chacon, P. (2009) FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics*, **25**, 2544-2551.
216. Ritchie, D.W. and Kemp, G.J. (2000) Protein docking using spherical polar Fourier correlations. *Proteins*, **39**, 178-194.
217. Vajda, S. and Kozakov, D. (2009) Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol*, **19**, 164-170.
218. Shen, Y., Paschalidis, I., Vakili, P. and Vajda, S. (2008) Protein docking by the underestimation of free energy funnels in the space of encounter complexes. *PLoS Comput Biol*, **4**, e1000191.
219. Pierce, B. and Weng, Z. (2007) ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins*, **67**, 1078-1086.

220. Feliu, E., Aloy, P. and Oliva, B. (2011) On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. *Protein Sci*.
221. Moont, G., Gabb, H.A. and Sternberg, M.J. (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**, 364-373.
222. Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V.A., Pieper, U., Stuart, A.C., Marti-Renom, M.A., Madhusudhan, M.S., Yerkovich, B. *et al.* (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res*, **31**, 3375-3380.
223. Davis, F.P. and Sali, A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901-1907.
224. Davis, F.P., Braberg, H., Shen, M.Y., Pieper, U., Sali, A. and Madhusudhan, M.S. (2006) Protein complex compositions predicted by structural similarity. *Nucleic Acids Res*, **34**, 2943-2952.
225. Lu, L., Lu, H. and Skolnick, J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350-364.
226. Chen, H. and Skolnick, J. (2008) M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys J*, **94**, 918-928.
227. Fleishman, S.J., Corn, J.E., Strauch, E.M., Whitehead, T.A., Andre, I., Thompson, J., Havranek, J.J., Das, R., Bradley, P. and Baker, D. (2010) Rosetta in CAPRI rounds 13-19. *Proteins*, **78**, 3212-3218.
228. Aloy, P. and Russell, R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161-162.
229. Yang, J.S., Campagna, A., Delgado, J., Vanhee, P., Serrano, L. and Kiel, C. (2012) SAPIN: Structural Analysis for Protein Interaction Networks. *Bioinformatics*.
230. Lasker, K., Sali, A. and Wolfson, H.J. (2010) Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins*, **78**, 3205-3211.
231. Lasker, K., Topf, M., Sali, A. and Wolfson, H.J. (2009) Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J Mol Biol*, **388**, 180-194.
232. Lasker, K., Phillips, J.L., Russel, D., Velazquez-Muriel, J., Schneidman-Duhovny, D., Tjioe, E., Webb, B., Schlessinger, A. and Sali, A. (2010) Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol Cell Proteomics*, **9**, 1689-1702.
233. Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B. and Sali, A. (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol*, **10**, e1001244.
234. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.S. *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*, **6**, 91-97.

235. Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M. *et al.* (2009) Literature-curated protein interaction datasets. *Nat Methods*, **6**, 39-46.
236. Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, **31**, 248-250.
237. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, **38**, D525-531.
238. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, **32**, D449-451.
239. Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*, **39**, D698-704.
240. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A. *et al.* (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res*, **37**, D767-772.
241. Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E. *et al.* (2011) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*, **40**, D857-861.
242. Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W. and Stumpflen, V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, **34**, D436-441.
243. Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.W. *et al.* (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832-834.
244. Han, K., Park, B., Kim, H., Hong, J. and Park, J. (2004) HPID: the Human Protein Interaction Database. *Bioinformatics*, **20**, 2466-2470.
245. McDowall, M.D., Scott, M.S. and Barton, G.J. (2009) PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res*, **37**, D651-656.
246. Brown, K.R. and Jurisica, I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076-2082.
247. Huang, T.W., Tien, A.C., Huang, W.S., Lee, Y.C., Peng, C.L., Tseng, H.H., Kao, C.Y. and Huang, C.Y. (2004) POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, **20**, 3273-3276.
248. Yellaboina, S., Tasneem, A., Zaykin, D.V., Raghavachari, B. and Jothi, R. (2011) DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res*, **39**, D730-735.
249. Gong, S., Yoon, G., Jang, I., Bolser, D., Dafas, P., Schroeder, M., Choi, H., Cho, Y., Han, K., Lee, S. *et al.* (2005) PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics*, **21**, 2541-2543.

250. Segura, J. and Fernandez-Fuentes, N. (2011) PCRPi-DB: a database of computationally annotated hot spots in protein interfaces. *Nucleic Acids Res*, **39**, D755-760.
251. Cukuroglu, E., Gursoy, A. and Keskin, O. (2011) HotRegion: a database of predicted hot spot clusters. *Nucleic Acids Res*, **40**, D829-833.
252. Guney, E., Tuncbag, N., Keskin, O. and Gursoy, A. (2008) HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res*, **36**, D662-666.
253. Thorn, K.S. and Bogan, A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284-285.
254. Winter, C., Henschel, A., Kim, W.K. and Schroeder, M. (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res*, **34**, D310-314.
255. Teyra, J., Doms, A., Schroeder, M. and Pisabarro, M.T. (2006) SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics*, **7**, 104.
256. Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J., Bolser, D.M., Oh, D., Kim, D.S. and Bhak, J. (2005) A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, **6**, 207.
257. Tuncbag, N., Gursoy, A., Guney, E., Nussinov, R. and Keskin, O. (2008) Architectures and functional coverage of protein-protein interfaces. *J Mol Biol*, **381**, 785-802.
258. Laskowski, R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res*, **29**, 221-222.
259. Hwang, H., Vreven, T., Janin, J. and Weng, Z. (2010) Protein-protein docking benchmark version 4.0. *Proteins*, **78**, 3111-3114.
260. Faure, G., Andreani, J. and Guerois, R. (2011) InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res*, **40**, D847-856.
261. Dreze, M., Monachello, D., Lurin, C., Cusick, M.E., Hill, D.E., Vidal, M. and Braun, P. (2010) High-quality binary interactome mapping. *Methods Enzymol*, **470**, 281-315.
262. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449-453.
263. Ben-Hur, A. and Noble, W.S. (2006) Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7 Suppl 1**, S2.
264. Jang, W.H., Jung, S.H. and Han, D.S. (2012) A Computational Model for Predicting Protein Interactions based on Multi-Domain Collaboration. *IEEE/ACM Trans Comput Biol Bioinform*.
265. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. and Jiang, H. (2007) Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*, **104**, 4337-4341.

266. Yu, C.Y., Chou, L.C. and Chang, D.T. (2010) Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics*, **11**, 167.
267. Lees, J.G., Heriche, J.K., Morilla, I., Ranea, J.A. and Orengo, C.A. (2011) Systematic computational prediction of protein interaction networks. *Phys Biol*, **8**, 035008.
268. Hart, G.T., Ramani, A.K. and Marcotte, E.M. (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol*, **7**, 120.
269. Trabuco, L.G., Betts, M.J. and Russell, R.B. (2012) Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods*.
270. Chiang, T. and Scholtens, D. (2009) A general pipeline for quality and statistical assessment of protein interaction data using R and Bioconductor. *Nat Protoc*, **4**, 535-546.
271. Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D. *et al.* (2010) The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*, **38**, D540-544.
272. Aragues, R., Jaeggi, D. and Oliva, B. (2006) PIANA: protein interactions and network analysis. *Bioinformatics*, **22**, 1015-1017.
273. Danielson, M.L. and Lill, M.A. (2010) New computational method for prediction of interacting protein loop regions. *Proteins*, **78**, 1748-1759.
274. McCammon, J.A. (1998) Theory of biomolecular recognition. *Curr Opin Struct Biol*, **8**, 245-249.
275. Tsai, C.J., Kumar, S., Ma, B. and Nussinov, R. (1999) Folding funnels, binding funnels, and protein function. *Protein Sci*, **8**, 1181-1190.
276. Levinthal, C. (1968) Are there pathways for protein folding? *J Chem Phys*, 44-45.
277. Tompa, P. and Rose, G.D. (2011) The Levinthal paradox of the interactome. *Protein Sci*, **20**, 2074-2079.
278. Park, Y. and Marcotte, E.M. (2011) Revisiting the negative example sampling problem for predicting protein-protein interactions. *Bioinformatics*, **27**, 3024-3028.
279. Park, Y. (2009) Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences. *BMC Bioinformatics*, **10**, 419.
280. Bjorkholm, P. and Sonnhammer, E.L. (2009) Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics*, **25**, 3020-3025.
281. Lee, H., Deng, M., Sun, F. and Chen, T. (2006) An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, **7**, 269.
282. Maetschke, S.R., Simonsen, M., Davis, M.J. and Ragan, M.A. (2012) Gene Ontology-driven inference of protein-protein interactions using inducers. *Bioinformatics*, **28**, 69-75.

283. Betel, D., Breitkreuz, K.E., Isserlin, R., Dewar-Darch, D., Tyers, M. and Hogue, C.W. (2007) Structure-templated predictions of novel protein interactions from sequence information. *PLoS Comput Biol*, **3**, 1783-1789.
284. Yu, J., Guo, M., Needham, C.J., Huang, Y., Cai, L. and Westhead, D.R. (2010) Simple sequence-based kernels do not predict protein-protein interactions. *Bioinformatics*, **26**, 2610-2614.
285. Aragues, R., Sander, C. and Oliva, B. (2008) Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics*, **9**, 172.
286. Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M. and Roth, F.P. (2009) Next generation software for functional trend analysis. *Bioinformatics*, **25**, 3043-3044.
287. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res*, **40**, D290-301.
288. Keskin, O., Nussinov, R. and Gursoy, A. (2008) PRISM: protein-protein interaction prediction by structural matching. *Methods Mol Biol*, **484**, 505-521.