Structural analysis of protein interaction networks

Anne Campagna

TESI DOCTORAL UPF / ANY 2011

DIRECTORS DE LA TESI
Prof. Dr. Luis Serrano & Dr. Christina Kiel
EMBL-CRG Systems Biology Unit
CRG-Center for Genomic Regulation

ABSTRACT

Interactions between proteins give rise to many functions in cells. In the last decade, high-throughput experiments have identified thousands of protein interactions, which are often represented together as large protein interaction networks. However, the classical way of representing interaction networks, as nodes and edges, is too limited to take dynamic properties such as compatible and mutually exclusive interactions into account. In this work, we study protein interaction networks using structural information. More specifically, the analysis of protein interfaces in three-dimensional protein structures enables us to identify which interfaces are compatible and which are not. Based on this principle, we have implemented a method, which aims at the analysis of protein interaction networks from a structural point of view by (1) predicting possible binary interactions for proteins that have been found in complex experimentally and (2) identifying possible mutually exclusive and compatible complexes. We validated our method by using positive and negative reference sets from literature and set up an assay to benchmark the identification of compatible and mutually exclusive structural interactions. In addition, we reconstructed the protein interaction network associated with the G protein-coupled receptor Rhodopsin and defined related functional sub-modules by combining interaction data with structural analysis of the network. Besides its established role in vision, our results suggest that Rhodopsin triggers two additional signaling pathways towards (1) cytoskeleton dynamics and (2) vesicular trafficking.

RESUMEN

Las funciones de la proteínas resultan de la manera con la que interaccionan entre ellas. Los experimentos de alto rendimiento han permitido identificar miles de interacciones de proteínas que forman parte de redes grandes y complejas. En esta tesis, utilizamos la información de estructuras de proteínas para estudiar las redes de interacciones de proteínas. Con esta información, se puede entender como las proteínas interaccionan al nivel molecular y con este conocimiento se puede identificar las interacciones que pueden ocurrir al mismo tiempo de las que están incompatibles. En base a este principio, hemos desarrollado un método que permite estudiar las redes de interacciones de proteínas con un punto de vista más dinámico de lo que ofrecen clásicamente. Además, al combinar este método con minería de la literatura y los datos de la proteómica hemos construido la red de interacciones de proteínas asociada con la Rodopsina, un receptor acoplado a proteínas G y hemos identificado sus sub-módulos funcionales. Estos análisis surgieron una novel vía de señalización hacia la regulación del citoesqueleto y el tráfico vesicular por Rodopsina, además de su papel establecido en la visión.

PUBLICATIONS

Anne Campagna, Luis Serrano, and Christina Kiel. (2008) Shaping dots and lines: Adding modularity into protein interaction networks using structural information, *FEBS Letters*, **582**, 1231-6.

Christina Kiel*, Andreas Vogt*, Anne Campagna*, Andrew Chatr-aryamontri, Magdalena Swiatek-de Lange, Monika Beer, Sylvia Bolz, Andreas F. Mack, Norbert Kinkl, Gianni Cesareni, Luis Serrano, and Marius Ueffing. Structural and functional protein network analyses predict novel signaling functions for the G protein-coupled receptor rhodopsin. *Molecular Systems Biology*. Accepted.

* These authors contributed equally to this work

Anne Campagna, Peter Vanhee, Luis Serrano and Christina Kiel. SAPIN: Structural Analysis for Protein Interaction Networks. *Bioinformatics*. In preparation.

ACKNOWLEDGEMENTS

It would not have been possible to write this PhD dissertation without the contribution and support of a number of people. In this section I would like to express my sincere thanks for any kind of help I received.

I would like to acknowledge my supervisors Luis Serrano and Christina Kiel for giving my the opportunity to pursue my PhD at the CRG and for supporting and guiding my work throughout the years.

I am also grateful to the members of my thesis committee Ben Lehner, Baldomero Oliva and Cedric Notredame for critically reviewing my work and providing me with insightful comments.

I have been very fortunate to be surrounded by great people within the Serrano Lab. I would like to especially thank my friends and colleagues from the dry lab: Kiana, Marie, Erik, Javier and Peter. It has been a real pleasure to share these few square meters with you! I also acknowledge all the past and current Serrano lab members for contributing to the nice work atmosphere in the lab.

During my PhD, I also had the opportunity to learn and play bass guitar in an almost famous band, a.k.a. Elsass Lightning a.k.a. Chatte Plastique. Thank you Holger, Michi and Tobias for giving me the chance to play with you and to make me experience few moments of glory during our concerts.

Throughout the years, I got to meet many people who made me have great time and supported me at the different stages of my PhD. Thank you Almer, Andreia, Bernd, Camilla, Cedrik, Elena, Mike, Phil, Ronan, Sarah, Sonja, Tony!

Very special thanks go to my whole family. First, my parents who have unconditionally supported any of my decisions and always encouraged me to constantly

surpass myself. My brothers, Franck and Olivier, whose living experience abroad has inspired me to move to Barcelona in the first place. Despite the long distances between us, you always managed to make me feel loved.
Finally, I would like to thank the most important person, and whose contribution to this thesis is priceless: Erik. Thank you so much for your unlimited help and patience when most needed and for always believing in me especially in the difficult moment. Thanks for being such a great inspiration source for me. Ik hou van ye!

CONTENTS

CHAPTER

ONE

INTRODUCTION

Proteins achieve their function by binding to other molecules. Depending on their localization in the cell, they can bind other proteins, DNA or RNA molecules, metabolites, intracellular and extracellular ligands, small nucleotides or lipids. These interactions are highly specific and usually involve the formation of a set of non-covalent bonds (*i.e.* Van der Waals, electrostatic and hydrophobic interactions).
Protein interactions can be stable or transient, specific or nonspecific, form homo- and hetero-oligomers, binary or part of larger protein assemblies. At the molecular level, protein interaction interfaces can involve either globular domains or short linear motifs.
The knowledge of protein interactions is crucial to understand how a cell works and interacts with its environment.

1.1 Large-scale analysis of protein interactions

1.1.1 Understanding protein interactions through Systems Biology

The technical progress made in genome sequencing and high-throughput experiments has initiated a major paradigm shift in biomedical research. Before the development of these "-omics" methods, molecular biologists were investigating individual genes and proteins belonging to complex biological processes. In the

past decade, the enormous amount of data that has been (and is still being) generated has opened the exciting possibility of reaching a better understanding of the various systems that are being studied for decades. However, if these approaches changed the way one looks at these systems, they have not made them less complex. And the famous statement made by Aristotle in the ancient times : "*The whole is more than the sum of its parts*", summarizes well the challenges and current focus of the system-wide approaches. Indeed, it has become clear that knowing the genes and proteins of a given cell is not enough to understand how a gene gets activated in response to a signal received by the cell. Similarly, knowing the mutations involved in a disease offers only partial information of the whole disease mechanism while understanding these mechanisms as a whole is crucial to treat diseases.
The protein interaction maps generated by large-scale analysis, usually graphically represented by networks, have been the subject of many analyses with the objective of deriving the properties inherent to the various systems. In this section, the main experimental methods to detect protein-protein interactions (PPIs) are described. We also briefly review the main existing repositories and resources available.

1.1.2 Identifying protein interactions

Yeast Two-Hybrid assay

The principle of the yeast two-hybrid system (Y2H) is based on the fact that many eukaryotic transcription activators have at least two distinct domains, one that binds to a promoter DNA sequence (BD) and another that activates transcription (AD)(Figure 1.1). These domains are modular and can function independently. It has been shown that if BD and AD are dissociated, the transcription gets inactivated. The modular property of these domains makes their concerted function possible even if the two domains are not covalently bound to each other, as long as they are physically associated to each other (Fields & Song, 1989). This subtlety has been the basis for the development of the Y2H method: transcription would get initiated if the two domains get close in space via an interaction of proteins X and Y. In practical, protein X is fused to BD (bait) and protein Y is fused to AD (prey). These chimeric proteins are cloned into expression plasmids and then transfected into a yeast cell. If X and Y physically interact, the reporter gene is activated.

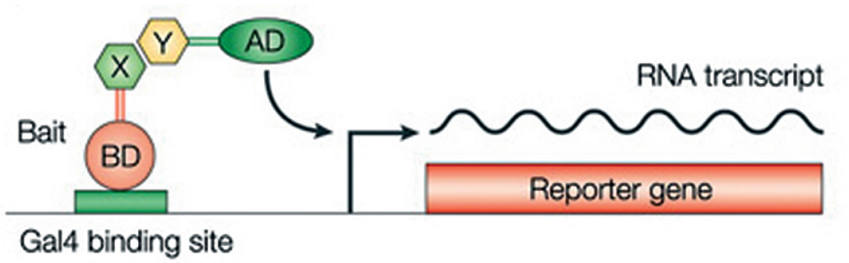


Figure 1.1: **Description of the Yest Two-Hybrid method.** - Figure from Grünenfelder & Winzeler (2002). The yeast two-hybrid technique measures protein-protein interactions by measuring transcription of a reporter gene. If protein X and protein Y interact, then their DNA-binding domain (BD) and activation domain (AD) will the transcription of the reporter gene.

However, this method can generate false positives when the transcription is activated but the interaction between bait and prey does not take place *in vitro*. Also, it can fail at detecting proteins that interact with each other (e.g. false negatives) if the fusion to BD or AD makes the interaction surface unavailable for binding with the tested interacting partners. Additionally, one has to take into account that a protein interaction can be detected by Y2H but, according to their sub-cellular localization, this interaction may not occur *in vivo*. One advantage of Y2H assays is they can identify interactions mediated by post-translational modifications. Moreover, these assays can detect many interactions in parallel. This method has indeed been applied to whole organisms since 2000 (Uetz *et al.*, 2000; Ito *et al.*, 2001; Stelzl *et al.*, 2005; Rual *et al.*, 2005; Giot *et al.*, 2003), by systematically testing all pairwise combinations of all the predicted proteins of an organism to derive the "binary" interactome. However, it has been pointed out that the unprecedented amount of data generated by these various large-scale Y2H experiments show a small overlap (Parrish *et al.*, 2006). This can be explained by different factors: the differences in protein interaction sampling, Y2H bias towards nonspecific interactions (Deeds *et al.*, 2006), and limitations of the Y2H method itself e.g. the transcription factors that cannot be targeted by this method and the potential change of protein structure conformations due to sequence chimera, which impedes a proper folding or interaction.

Tandem affinity purification method

A Tandem Affinity Purification (TAP) tag (Rigaut *et al.*, 1999) involves the fusion of two affinity modules to the protein of interest and the introduction of the construct into the host cell or organism (Figure 1.2). In a first affinity step, the first tag (*e.g.*, protein A) binds to a first affinity column (*e.g.*, IgG beads). The contaminant proteins coming from the cell extract are then washed out and a protease cleaves the link between the two tags. A second affinity purification involves for example calmodulin-coated beads that are incubated together with the eluate in presence of calcium. The target protein complex is finally released. The fragments resulting from the cleavage of the complex components by proteases are identified by Mass Spectrometry (MS). The advantage of using two tags significantly reduces non-specific background, as compared to a single tag approach. Many proteins can be associated with the bait protein, but this method does not provide any information about the arrangement of the complex components. In contrast to the Y2H method, the TAP tag approach allows the identification of *in vivo* protein complexes. The interactions identified by TAP tag are more likely part of large and stable molecular machines rather than transient interactions which are less stable and consequently leads to the proteins to be washed away. In addition, the identified complexes could be biased towards abundant proteins (von Mering *et al.*, 2002). This method has also been applied in a large-scale fashion in yeast (Gavin *et al.*, 2002; Krogan *et al.*, 2006) and more recently in human (Ewing *et al.*, 2007), to generate a "co-complex" interactome. A big advantage of this approach is that it preserves the biological context of the complexes. However, because of the large number of proteins that can be found to be associated with a given bait protein, it is most likely that more than one complex is associated with a given bait protein. Additionally, the overexpression of the bait can result in the association of the bait protein with chaperones and can lead to improper intracellular localization.

The data generated by TAP tag approaches usually requires extra analyses in order to determine which of the identified components of a complex are physically interacting with each other. For this purpose, two models are commonly used to elucidate a complex topology: the 'spoke' model and the 'matrix' model (Bader & Hogue, 2002). The spoke model supposes that the bait protein interacts directly with each

component of the complex, excluding a physical association between any of the latter. The matrix model assumes that all the proteins identified within a complex form binary interactions with each other. To reach a more functional understanding of the co-purified protein associations, a socio-affinity index measuring the propensity of proteins to form partnerships has been further developped by Gavin *et al.* (2006).

Although Y2H and TAP tag approaches are complementary, it is important to note

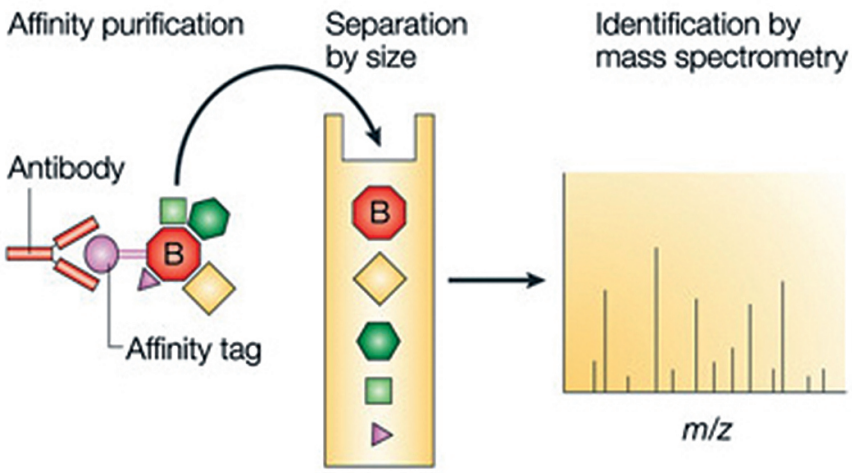


Figure 1.2: **Description of the Tandem Affinity Purification method, followed by Mass Spectrometry.** - Figure from Grünenfelder & Winzeler (2002). Affinity Purification methods can isolate a protein complex associated to a tagged protein. The components of the complex are ths separated and identified by mass spectrometry.

their differences, especially when it comes to define what "protein-protein interaction" refers to. The Y2H method identifies binary physical interactions while a TAP tag approach provides a list of the components belonging to one or more complexes. There has been confusion among the scientific community partly because in co-complex experiments, the term"interaction" has been used to describe the relationships between proteins in a complex and not binary direct interactions (Gavin *et al.*, 2002; Krogan *et al.*, 2006).

Resources for protein interaction data

The recent large-scale determination of protein interaction data has led to the development of databases in an attempt to organize, classify and easily access this information. Table 1.1 provides an overview of the available resources. Some of these databases contain data coming exclusively from experiments, while others combine experimental and predicted data. In addition to encompassing the main large scale experiments that have aimed at providing a nearly complete catalogue of the interactions for many organisms (Uetz *et al.*, 2000; Ito *et al.*, 2001; Gavin *et al.*, 2002; Giot *et al.*, 2003; Rual *et al.*, 2005; Stelzl *et al.*, 2005; Krogan *et al.*, 2006), much effort has been made for implementing literature mining tools to include previously published interaction data. However, these databases have been subject to debate about the quality of the data (Mackay *et al.*, 2007; Chatr-Aryamontri *et al.*, 2008). In fact, it has been pointed out that many proteins identified in co-complex experiments were annotated as pairwise protein interactions. Moreover, automated literature mining methods may associate by mistake two proteins based on wrongly annotated names. In research articles, authors usually use gene or protein names which are different from the name used in databases. To facilitate the integration of diverse data from millions of articles, and increase the quality of interaction data annotation, Ceol *et al.* (2008) have suggested the use of structured abstracts where authors would provide their findings in an organized way, making it easy to integrate in the existing resources.
In addition, many of these databases provide a score that typically takes into account the number, type and size of experimental evidence. In these calculations, data coming from high-throughput experiments are usually assigned a low weight while interactions which have been observed multiple times get a higher score.

1.1.3 Global properties of protein interactions networks

A logical first approach to analyze protein interaction networks has been to describe their global properties. Since the sets of interactions are usually represented as nodes connected by edges in a graph, the first network analyses have naturally borrowed characteristics traditionally used in the field of network theory. The ubiq-

Database	Type of data	Reference
APID	Meta	Hernandez-Toro *et al.* (2007)
BIND	Exp	Bader *et al.* (2003)
BioGRID	Exp	Stark *et al.* (2011)
DIP	Exp	Xenarios *et al.* (2002)
HPRD	Exp	Keshava Prasad *et al.* (2009)
IntAct	Exp	Aranda *et al.* (2010)
MiMI	Exp,Pred	Jayapandian *et al.* (2007)
MINT	Exp	Ceol *et al.* (2010)
MIPS-MPact	Exp	Güldener *et al.* (2006)
MIPS-MPPI	Exp	Pagel *et al.* (2005)
MPIDB	Meta	Goll *et al.* (2008)
OPHID	Exp,Pred	Brown & Jurisica (2005)
PINA	Meta	Wu *et al.* (2009)
PIPs	Exp,Pred	McDowall *et al.* (2009)
STRING	Meta	Szklarczyk *et al.* (2011a)
UniHI	Exp,Pred	Chaurasia *et al.* (2007)

Table 1.1: **Main resources for Protein interaction data** - The type of data indicates if the database contains experimentally identified interactions (Exp) or predicted interactions (Pred), or if the database is a meta-database (Meta).

uitous identification of scale-free networks described by Barabási (1999), based on their analysis of the World Wide Web, and the simultaneous increase of large scale PPI data have initiated the interest in topological properties of biological networks. Since then, numerous studies showed that PPI networks often exhibit a scale-free behavior, which means that they have many nodes with small degrees (e.g. with few connections) and allow nodes with high degrees (e.g. with many connections) with decreasing probability. Network topologies have then been linked to a variety of biological implications. For example, it has been shown that the most highly connected proteins in the cell, also called "hubs", are essential for its survival (Jeong *et al.*, 2001). Also, an important property of scale-free networks is their robustness: if the network is disturbed by random events (e.g. node failures) there is a high probability that an essential node will not be affected and therefore the network could still be functional. In addition, the topology of PPI networks varies according to the type of experiment producing the data. Figure 1.3 shows that a "binary" inter-

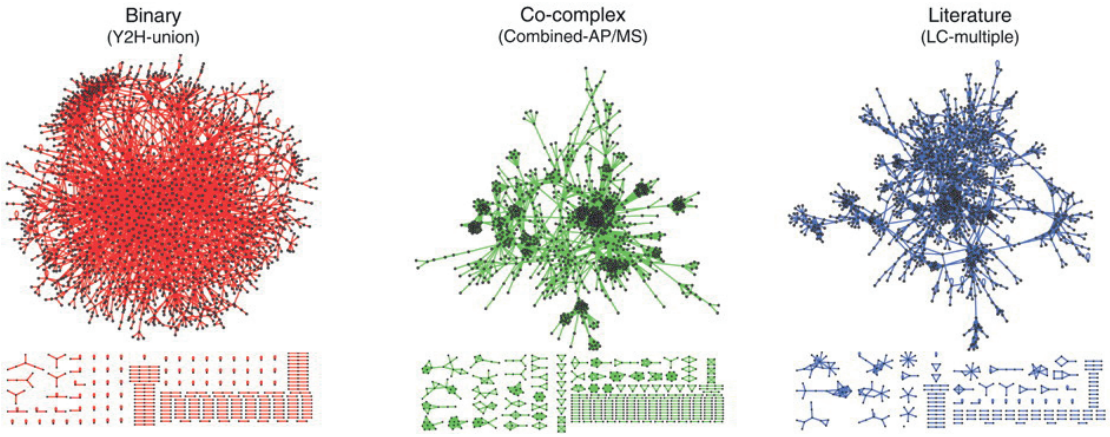


Figure 1.3: **Network representation of interaction data** - Figure from Yu *et al.* (2008). The protein interactions networks generated by Yeast Two-Hybrid assays (left), Affinity Purification followed by Mass Spectrometry (middle) or from literacture data (right) exhibit different topologies.

actome coming from Y2H assays contains few hubs in contrast with a "co-complex" interactome generated by TAP tag experiments.

In addition to large scale experiments exploring all putative protein-protein interactions in a cell (Uetz *et al.*, 2000; Ito *et al.*, 2001; Stelzl *et al.*, 2005; Rual *et al.*, 2005; Giot *et al.*, 2003; Gavin *et al.*, 2002; Krogan *et al.*, 2006), computational tools have been developed to predict protein interactions. The objective of these methods is to increase the usually low coverage observed in the experimental interaction datasets but also to infer physical interactions from the protein complexes. Some of these predictors are based on comparative genomics (Pellegrini *et al.*, 1999; Enright *et al.*, 1999; Goh *et al.*, 2000; Huynen *et al.*, 2000), others infer the interactions using orthology (Lehner & Fraser, 2004) or are based on functional features, e.g the expression levels of transcripts encoding the proteins (Jansen *et al.*, 2003) and sub-cellular localization (Deane, 2002; Gandhi *et al.*, 2006). Finally, another type of predictors exploits the observation that some pairs of sequence motifs, domains and structural families tend to interact preferentially.

The large amount of interaction data produced by high-throughput experiments and data-mining approaches enabled the construction of PPI networks. This way

of representation is convenient to analyze global properties of a system but they depict a rather static picture of highly dynamical and regulated processes.

1.2 Protein interactions from a structural point of view

1.2.1 The role of structural data for the study of protein interactions

Following genomics projects, post-genomics projects have aimed at providing the full list of proteins and their interactions within an organism. The tremendous amount of data they produced gives insights about the global properties of biological systems. However, represented together in PPI networks, it provides a rather static image of highly dynamical events taking place in the cell.
The function of a protein often relies on the way it interacts with its partners. In general, protein-protein interactions involve a proportionally small surface compared to the number of its interacting partners. As a result, a protein could interact with many others using different surfaces, but also in many cases, two proteins could compete for the same surface. Structural data can thus help improving our understanding of complex systems, by adding the important missing feature of competition into protein interaction networks. If proteins compete, they cannot bind at the same time and therefore the time differences regarding when they are expressed, or their localization is essential to know which interactions will take place.
The characterization of "hubs" within PPI networks shows that proteins can have a large number of partners and it is therefore obvious that they cannot simultaneously physically interact with this "hub". The knowledge of 3D structures can be used to identify the interactions that can occur simultaneously and the ones that are mutually exclusive.
The characterization of competing interfaces at the structural level was first explored by Kim *et al.* (2006a). They used structures of domain interactions to investigate whether hubs interact using many interfaces (multi-interface hubs) or a single interface (singlish-interface hubs). They observed that multi-interface hubs evolve

more slowly than the others. Additionally, the multi-interface hubs were found to be more closely correlated with their partner expression levels compared with singlish-interface hubs. These findings provided a structural explanation of expression dynamics for hubs and their interacting partners.
The concept of time dimensionality in protein interaction networks has also been applied by Tuncbag *et al.* (2009) on the p53 network. They built a structure-based network of p53 and, by analyzing the protein interactions that could co-exist, they predicted four distinct binding sites on the DNA-binding domain, involving 12 interacting proteins (Figure 1.4).

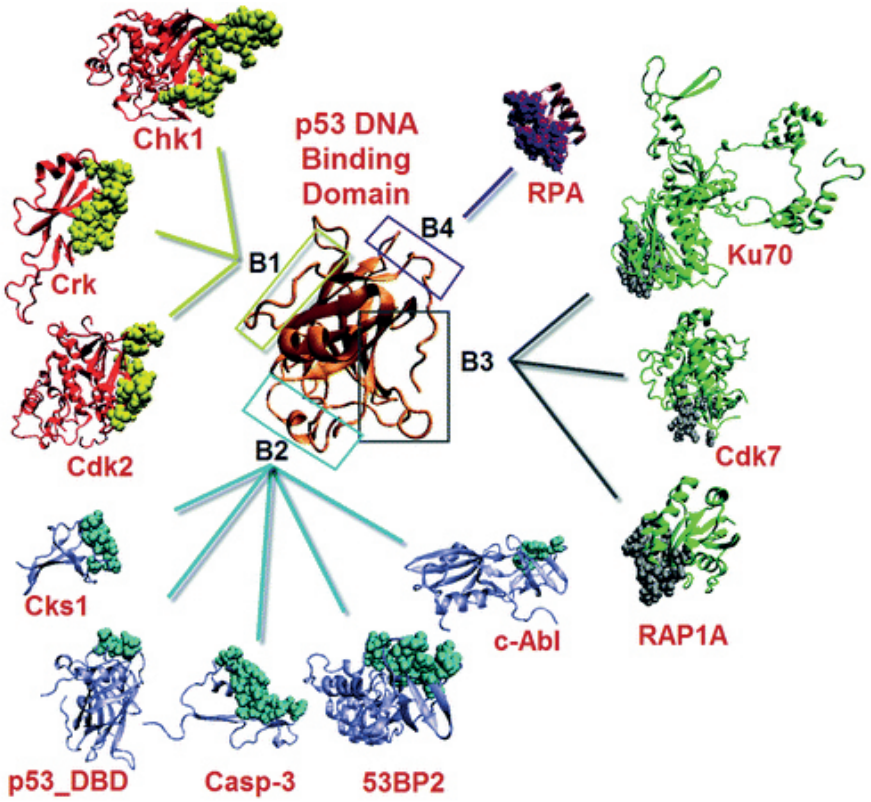


Figure 1.4: **Predicted partners for the p53 DNA binding domain.** - Taken form Tuncbag *et al.* (2009). The p53 DNA Binding Domain (orange, center) engaged interactions through four interfaces (B1,B2,B3,B4). The structural information makes possible the identification of competing binding for three of the surfaces.

These studies illustrate well the importance of using structural information at

the network level, since it has the ability to convert two-dimensional networks into functional pathways and can help gain insight into the understanding of dynamical and regulated processes.

In addition, knowing the interfaces through which proteins interact allows to derive the molecular phenotypes associated with disease mutations: if these are located at the protein surface, they are more likely to affect a specific interaction by either decreasing the binding affinity or completely impeding the binding, while the interactions occurring through another part of the surface can remain unaffected. Recently, Zhong *et al.* (2009) investigated the effect of mutations at the network level, by mapping 3664 mutations at the surface of 249 protein structures and they suggested two distinct network perturbations for human inherited disorders: "node removal", which results from the loss of gene products, and "edgetic perturbation", altering specific molecular interactions. Following the same principle, the structural knowledge of protein interactions can also serve as a basis for computational protein design, if one wants to mutate specific residues at a protein surface to selectively disable a downstream pathway.

Given three proteins A, B and C, they can interact following three possible conformations (Figure 1.5):

1. **A-B-C-A**: the three proteins form binary interactions (Figure 1.5, top row).
2. **B-A-C**: B and C physically interacting with A through a different interface (Figure 1.5, middle row).
3. **A-B**/**A-C**: B and C physically interacting with A through the same interface (Figure 1.5, bottom row).

The network representation derived from PPI detection methods for these three cases can lead to a misinterpretation of their actual relationships, as shown in columns 2 and 3 of Figure 1.5. In fact, except if the proteins follow the type 1 of interactions, both binary and co-complex determination methods fail at reproducing the biological relationships between these three proteins. However, using structural information, one can reach the real topology of the connected proteins. This example, applied to the simplest case of interaction network, illustrates well

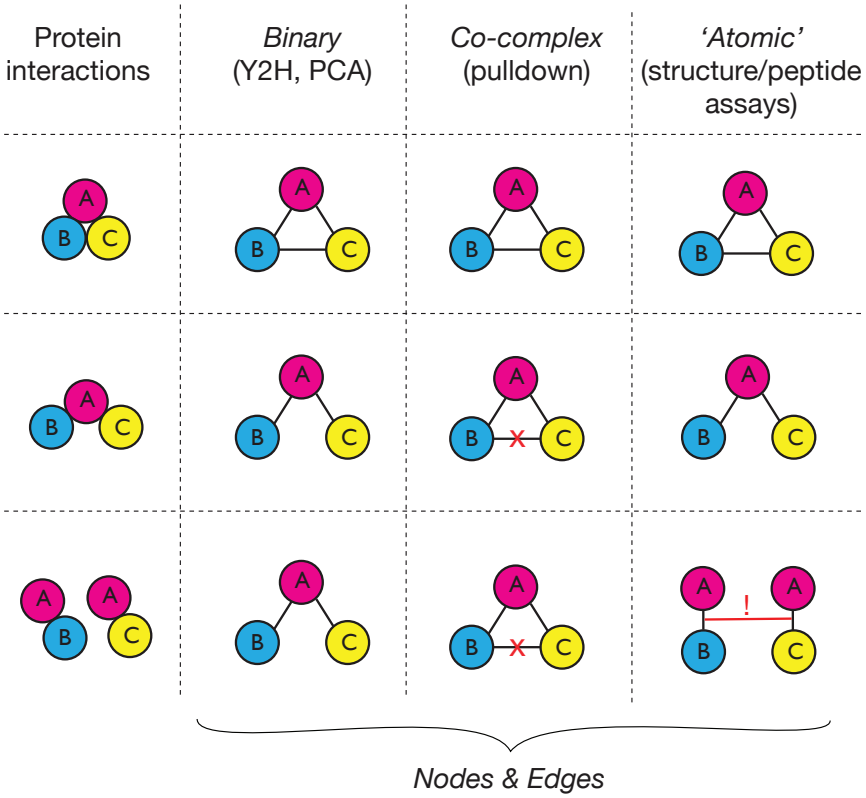


Figure 1.5: **Illustration of PPI networks derived from Binary, Co-Complex and Structural Determination methods.** - The left column illustrates the different ways of interacting for three given proteins A, B an C. The three columns on the right show the nodes-and-edges representation of these interactions according to Yeast Two-Hybrid, affinity purification and structural determination methods.

how integrating structural data into PPI networks can be of great value.

In the following sections, we briefly describe the experimental methods used to resolve protein structures and the computational approaches developed to cope with the low structural coverage.

1.2.2 Protein structure determination

The most common experimental methods to study tertiary protein structure are X-Ray crystallography and Nuclear Magnetic Resonance (NMR) spectrometry.

X-Ray crystallography

X-Ray crystallography is a very powerful method to analyze protein structures. It can provide the coordinates of the protein amino acids or DNA/RNA nucleotides at the atomic level. It determines the position of atoms within a crystal through exposing the crystal to a beam of X-rays. The atoms in the molecule cause the X-rays to diffract into many different directions, which results in a diffraction pattern that is recorded on photographic film. From the angles and intensities of these diffracted beams it is possible to mathematically construct the three dimensional image of the electron density in the crystal structure. Subsequently, from this electron density, the mean positions of the atoms in the crystal can be determined, as well as their chemical bonds and their disorder or order. All the resolved structures are gathered in the Protein Data Bank (PDB) (Rose *et al.*, 2011) and the coordinates and experimental procedures are described in a PDB file, which is typically a text file with specific fields that define the characteristics of the structure.
The growth of protein crystals of sufficient quality for structure determination is the rate-limiting step in most protein crystallographic work. To facilitate this process, proteins are often crystallized at high concentrations which may result in a non native conformation for these proteins. In addition, the tight packing of molecules can lead to the observation of contacts that are not related to the biological function of the proteins. In fact, a crystal may contain multiple copies of a protein or a protein complex so one should pay extra attention when working with this type of data. An example of crystal packing is shown in Figure 1.6. Many methods have attempted to identify these crystallographic artifacts by comparing the physical properties with those of biologically relevant interfaces (Levy, 2007; Zhu *et al.*, 2006; Bernauer *et al.*, 2008). The PDB itself has helped to cope with this issue by providing an extra field within each entry describing the parts of the structure that belong to the biological unit. In fact, the authors of the resolved structures have now provided this information. Being familiar with the system they work on, they often conducted additional biochemical experiments to characterize the biological unit.
X-Ray crystallography captures a rigid conformation of proteins, which impedes the observation of structurally variable parts of proteins, such as loops, which can change conformation according to their binding partners. Nevertheless, it is gen-

erally accepted that this technique provides protein structures with the highest resolution.

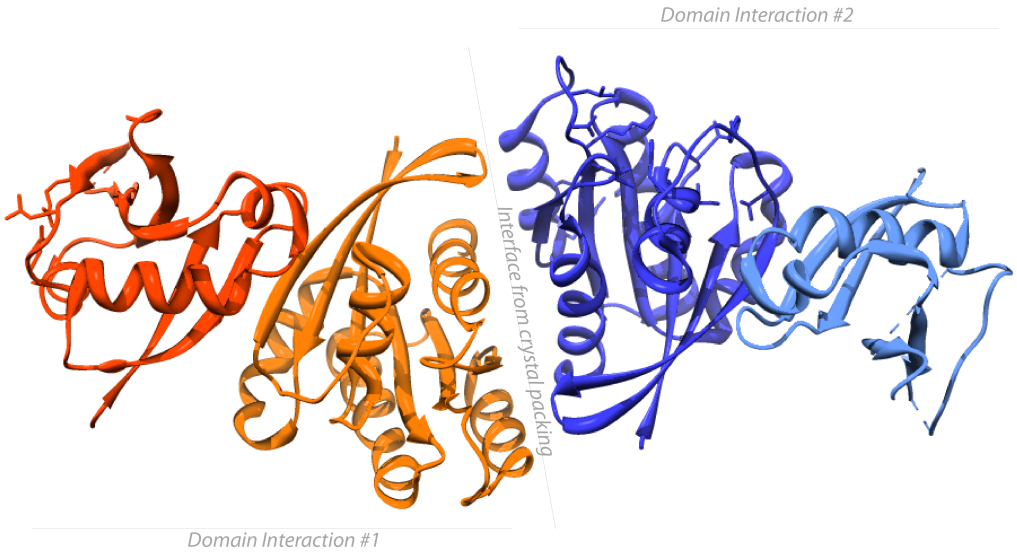


Figure 1.6: **Illustration of crystal packing** - The interface between the orange and the blue domain is a crystallographic artifact while the interface between the orange and the red domain is biologically relevant.

Nuclear Magnetic Resonance spectrometry

Nuclear Magnetic Resonance (NMR) spectroscopy can determine the three-dimensional structure of small proteins in aqueous solution. Basically, a sample gets placed in a magnetic field so that the spins of its protons are aligned. When radio-frequency pulses are applied, the protons get excited and emit signals, whose frequency depends on the molecular environment of the proton. From these signals it is possible to derive the protein structure. In aqueous solution, protein structures are more

flexible than in a crystal lattice. As a consequence, this conformational freedom results in an ensemble of NMR models. NMR spectroscopy often provides structures with lower resolution than X-Ray crystallography.

The clear limitation of these approaches as a protein interaction detection method remains the limited structural coverage at the network level. In human, less than 10% of the protein interactions have structural data (Figure 1.7). This low coverage has encouraged the development of computational tools aiming at bridging the gap between the low amount of structures available and the large number of experimentally determined interactions.

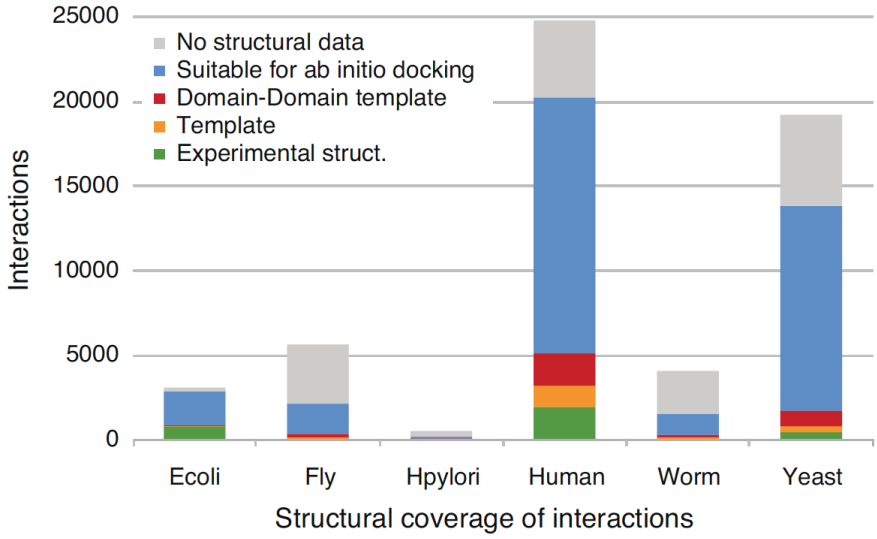


Figure 1.7: **Structural coverage of protein interactions.** - Taken form Stein *et al.* (2011b). For human and yeast, the use of domain-domain structural templates (red) can increase the structural coverage (green).

1.2.3 Protein interactions and modularity

Proteins are the most structurally complex and functionally sophisticated molecules known. They execute nearly all the cell's functions and are present in all forms of life. It is clear that the key for understanding protein function relies on the careful analysis of its structure. Depending on its conformation, a protein can uncover parts of its surface that can engage in the interaction with another protein. This could consist in the rearrangement of its globular domains or a loop moving out of the rest of the globular structure. Before analyzing the mechanisms of interactions, we take a closer look at the features that confer a protein its function.

Domains are the functional units of proteins

Protein domains are regarded as functional units of a protein. They are able to fold and evolve independently from the rest of the polypeptide chain they belong to (Ponting & Russell, 2002). If a protein consists of several domains, its function may be achieved either through one the individual domains or by a combination of domains, relying on their spatial arrangement. Domains can have structural roles if they are part of large assemblies or play a catalytic and/or regulatory role. The fact that domains have the ability to fold independently from the rest of the protein has permitted molecular and functional studies to focus on these subunits rather than full-length proteins. This is of great value to cope with technical limitations that are related to handling large proteins. Throughout evolution, domains have been duplicated, recombined, inserted and depleted within proteins, leading to new proteins with novel functions (Orengo & Thornton, 2005).

There are many domain resources available. Some are based on the analysis of repeatedly occurring folds while other resources are based on the sequences derived from observed and classified folds. In the first category of domain resources, the Structural Classification Of Proteins (SCOP) provides a detailed and comprehensive description of the structural evolutionary relationships of proteins. The classification hierarchically groups structures by fold, class, superfamily and family (Murzin *et al.*, 1995; Andreeva *et al.*, 2004). Similarly, the CATH database (Orengo *et al.*, 1997; Greene *et al.*, 2007) classifies the structures according to the follow-

ing hierarchy: Class, Architecture, Topology and Homologous Superfamily. These classifications are the results of the combination of manual and automated procedures, which allows the comparison and/or clustering of structures even if they exhibit low sequence similarity. The second category of domain resources provides a sequence-based definition of domains. PFAM (Sonnhammer *et al.*, 1997; Finn *et al.*, 2010a) and SMART (Schultz *et al.*, 1998a; Letunic *et al.*, 2009) are among the largest protein domain databases. Their families are derived from multiple sequence alignments from which Hidden Markov Models (HMMs) are built and can be used to search matches to domain families in sequence databases. The increasing number of sequences generated within the last decade by genomics projects has permitted the domain family definition to be improved and refined.

At the protein interaction network level, the organization in domains confers proteins an important modular property, which has been found to play an important role in regulatory and signal transduction pathways (Pawson & Nash, 2003; Seet *et al.*, 2006). Additionally, the fact that they are often engaged in transient and stable interactions with other domains or short linear motifs is an important feature for these types of pathways.

Domain-domain interactions

The structures of domain-domain interactions (DDIs) have been intensively studied, according to their type (interchain or intrachain), their topology, their interface and their binding sites. The resulting classifications are available in databases, such as iPfam (Finn *et al.*, 2005), 3did (Stein *et al.*, 2009), SCOPPI (Winter *et al.*, 2006), PRISM (Ogmen *et al.*, 2005), SNAPPI-DB (Jefferson *et al.*, 2007), and PIBASE (Davis & Sali, 2005). However, this information about structural DDIs only gives a hint about the availability of 3D data for a given pair of domain. The classification of the domain interaction interfaces by Kim *et al.* (2006b) has shown that about 60% of the domain family pairs have a single interaction topology, whereas the remaining 40% show multiple ways of interacting. These results indicate that focusing on the domains alone is not sufficient to elucidate the molecular mechanism of interac-

tions. The use of extra information about the domain families or the analysis of their interface properties could help narrowing down the number of possible interfaces and finally provide a more suitable template for the interaction.

Nevertheless, these resources represent a valuable tool to extend the structural knowledge of domain interactions to family members that lack a resolved structure. Aloy & Russell (2004) estimated the number of protein interactions to be limited to 10000 in Nature. In another study, they showed that pairs of interacting domains belonging to the same family tend to interact in a similar way (Aloy *et al.*, 2003). Taken together, these findings suggest that structural information of interacting domains can be exploited by computational approaches to provide relatively accurate models for protein interactions.

Domain-motifs interactions

Linear motifs or peptides are between 4 and 40 amino acids long and perform a wide range of functions both in cell-to-cell and intracellular communication: they are important mediators in many signaling pathways and regulatory networks (Neduva & Russell, 2006). In general, these short segments are characterized by local flexibility. Many of them, such as phosphorylation sites (Iakoucheva *et al.*, 2004), SH3 interaction motifs (Beltrao & Serrano, 2005) or recognition elements of 14-3-3 proteins (Bustos & Iglesias, 2006), have been found in locally disordered regions of their parent proteins. Known motifs are catalogued by several resources, including the eukaryotic linear motif (ELM) database (Gould *et al.*, 2010), SCANSITE (Obenauer, 2003) and PROSITE (Sigrist *et al.*, 2010). Many of the best-known motifs, such as those interacting with SH2 (Src Homology 2), PTB (phosphotyrosine binding) and 14-3-3 domains, recognize sites of post-translational modification made during signal transduction.

Post-translational modifications

Proteins are subjected to many reversible post-translational modifications (PTMs), such as the covalent addition of a phosphate or an acetyl group to a specific amino acid side chain. The addition of these modifying groups is used to regulate the activity of a protein, changing its conformation, its binding to other proteins and/or its location inside the cell. A striking example is the tumor suppressor protein p53: in its active form, p53 is subject to a complex and diverse array of covalent post-translational modifications, which influence the expression of p53 target genes (Toledo & Wahl, 2006). In fact, p53 is reported to bind to 380 proteins, according to the STRING database (Szklarczyk *et al.*, 2011b), and the specificity for each of them seems to be dependent on the many possible combinations of its various PTMs. The available resources for PTMs include Phospho.ELM (Dinkel *et al.*, 2011), netPhorest (Miller *et al.*, 2008) and SCANSITE (Obenauer, 2003).

1.2.4 Bridging the gap: using computational structural biology to model protein interactions

Structural information regarding proteins is very low compared to the sequence data that is being generated. This gap increases dramatically regarding protein interactions. Similarly to the numerous methods that attempt to predict the structure of a protein, the exponential increase of interaction data has encouraged the development of methods aiming at modeling protein interactions. A large number of predictors use observations and patterns derived from studying different protein interaction mechanisms at different levels of resolution (from low to high):

Protein docking

Docking procedures use surface complementarity and electrostatics to predict structural complexes from single structures of proteins, fitting together two or more known structures of reliable 3D structural models via their interacting surfaces (Katchalski-Katzir *et al.*, 1992). Docking programs usually comprise two steps: generation of thousands of alternative poses to sample all possible

interaction modes, followed by scoring these poses using a 'pseudo-energy' function. Approximately correct solutions are generated by the first step, but scoring functions often fail to rank them properly, as evaluated by the Critical Assessment of PRedicted Interactions (CAPRI,(Janin, 2005)). However, Wass *et al.* (2011) recently showed that a standard docking program can distinguish the true interactors from a background of 922 non-redundant potential interactors. Hence, despite the limitations of docking algorithms, they can still be useds to distinguish between binding and non-binding partners.

Comparative modeling-based method

The goal of comparative modeling - or homology modeling - is to build a structural model of a protein on the basis of close sequence similarity to a template protein of known structure. This principle can be extrapolated to the prediction of protein interactions, since it has been shown that homologous pairs of proteins tend to interact in the similar way (Aloy *et al.*, 2003). Aloy & Russell (2002) have exploited this property to develop a method based on the homology between interacting pairs. In addition, the contacts between the amino acids at the interface are being assessed using empirical potentials to determine the suitability of the template.

Protein interface modeling

Protein interface modeling restricts the focus to the residues directly involved in the interaction. For a given interaction, all the elements (i.e. secondary structure elements and loops) that are not involved in the complex formation are excluded from the template. The side chains of the remaining residues in the template are then substituted *in silico* by the side chains of the corresponding residues in the sequence of the proteins to be modeled. Kiel *et al.* (2007) used this approach to predict 20 Ras proteins in complex with 50 Ubiquitin-like domains with FoldX (Guerois *et al.*, 2002; Schymkowitz *et al.*, 2005), an all-atom empirical force-field (see Box). The resulting network showed very high accuracy for distinguish between binders and non-binders when compared to

pull down experiments. This methodology can be applied to the prediction of binding partners in other domain families, if enough structural information is available. However, the quality of the results depends on the quality of the structural templates and on correct structure-based sequence alignments. Thus this method requires a careful manual inspection of the structures and alignments and could not be applied in an automated process.

Estimation of the interaction energy using all-atom force fields
The stability of proteins and the strength of interacting proteins can be measured by diverse experimental methods but when working with structures of proteins instead of the actual protein one needs what is commonly called a 'force field' in computational structural biology. The 'FoldX' software is an example of such a force field that can be used for rapid evaluation of folding and binding energies or effects of mutations based on the three-dimensional coordinates of a structure (Schymkowitz *et al.*, 2005). FoldX makes a quantitative estimation of the stability of a structure or a complex by summing the positive and negative intramolecular forces (see table 1.2) and weighing their contribution based on empirical data from protein engineering experiments. The predictive power of the force field was tested on a large set of wild-type and mutated structures covering many different folds and environments (Guerois *et al.*, 2002).

1.3 From protein interaction networks to the "3D interactome" [1]

The techniques used to determine protein interactions typically identify binary interactions (i.e. Y2H) and protein complexes (i.e. TAP tag). The latter identify protein complexes whose components are not necessarily physically interacting with each other and the large number of protein co-purified with a bait suggests that more than one functional complex could be associated with this bait protein. The main advantage of these methods is that they can be applied in a high-throughput manner, leading to a large catalogue of interactions at the organism level. The representation by PPI networks of the data generated by these high-throughput approaches have enabled the characterization of "hub" proteins, which usually have more interacting partners than available surface for binding. Thus, it is obvious that the

1. the term 3D interactome has been introduced by Stein *et al.* (2011c)

Force type	Description
van der Waals interactions	Attractive or repulsive forces between adjacent molecules. The attractive forces cone from electron density fluctuations between adjacent uncharged non-bounded atoms. The repulsive forces are generated when the distance between two atoms is lower than the sum of their van der Waals radii.
Hydrogen bonds	Two electronegative groups ($N-$ or $C=O^{-}$) compete for the same hydrogen atom which is covalently attached to one of them (donor).
Hydrophobic interactions	Aggregation of non-polar compounds when surrounded by polar water molecules.
Electrostatic interactions	Long distance cohesive forces between negatively and positively charged atoms.
Entropy	Amount of disorder lost by forming bonds or restricting a conformation.

Table 1.2: **Non-covalent forces contributing to interaction energy**

interacting proteins will undergo competition for binding to this highly connected node. Stein *et al.* (2011c) recently evaluated the structural coverage for proteins and protein interactions for several organisms and showed that using 3D structures of domain interactions can increase the structural coverage to around 50 % for the human interactome. This number can be explained by the fact that domain folds are more conserved than their sequences. If one wants to model a protein, searching a potential template in the PDB for this protein sequence might not provide any structure suitable for modeling if they share less than 30% sequence similarity. The large number of sequence data used to generate HMM profiles in domain family databases has enabled the access to more remote members and thus possibly more potential structural templates.

This has encouraged the development of computational techniques to elucidate the molecular mechanisms of protein interactions. They range from docking to comparative modeling, with hybrid approaches falling in between (see above). However,

these techniques may be computationally expensive or require a detailed manual inspection of the data, which is a limiting factor to their application at the interactome level.

Structure-based classifications of domain interactions have demonstrated that a pair of interacting domains could exhibit multiple interaction topologies (Kim *et al.*, 2006b). In order to select the most appropriate structural template as a model for a protein interaction, it is essential to combine this data with an extra source of data, such as information about the interface.

To summarize, the knowledge of interacting interfaces provided by structural information can help determining the binary interactions taking place within experimentally characterized protein complexes and contribute to the elucidation of competition between binding proteins, by discriminating compatible and mutually exclusive interactions. We proposed to integrate this information in protein interaction networks by visually representing it using the logic gates "AND" (compatible) and "XOR" (mutually exclusive) (see Campagna *et al.* (2008), in section 1.4). Additionally, predicting interactions using DDI structures can help increasing the structural coverage. The sequence-based annotation of domains can additionally reduce the computation time for these predictions, which would thus make a domain-based structural approach to interaction modeling suitable to be automatically applicable to large protein interaction sets.

1.4 Appendix

Campagna A, Serrano L, Kiel C. Shaping dots and lines: adding modularity into protein interaction networks using structural information. FEBS Lett. 2008 Apr 9;582(8):1231-1236.

CHAPTER

TWO

OBJECTIVES

The main idea developped in this thesis was to propose structural information as very important to be combined with protein interaction networks. We first reasoned that if a central "hub" protein interacts with many partners using a similar structural interface, the binding proteins would compete with each other, if this hub is in sub-stoichiometric concentration expressed in a given biological system or cell type. Second, interaction data derived from Co-immunoprecipitation experiments are not necessarily all binary interactions, which means proteins could be found experimentally in complex without directly binding to the bait proteins; however, information on binary interactions combined with the knowledge of domains and linear motifs mediating the interaction and the interfaces that are used is crucial if we want to analyze the role of competition. In contrast to the field of *in silico* predictions of protein-protein interactions, which is limited by a low structural coverage and is highly error prone, we here propose something conceptually new: combining *in silico* predictions of domain and linear motif interactions with experimental information. We hypothesize that the prediction success could be improved if proteins were already found experimentally to be in one complex. Third, we wanted to analyze whether the structural and competition information, combined with protein interaction networks and signaling pathways could gain biological insights; therefore, we have tested this hypothesis on the rhodopsin signal transduction pathway.

1. **Rhodopsin signal transduction pathway as a test case to gain biological insights when combining experimental interaction data with structural**

information

Here, the objective was to combine structural information with Co-immunoprecipitation experiments that were performed in a defined cellular region: the rod outer segment (ROS). We aimed to get an estimate of the structural coverage, to develop a scoring system, and finally to find an objective way to establish an "exclusion criteria" to decide whether two proteins are compatible ("AND") or mutually exclusive ("XOR") for binding to a common third partner protein. Finally, we aimed to understand whether the information on AND and XOR interactions relates to the biological function of certain modules in the network.

2. **Developing a web server that combines structural information with networks, and proposing a way of visualizing the competing interactions**

   Here, the aim was to develop an automated pipeline that predicts domains and linear motifs that could mediate direct binary contacts within proteins that were found experimentally to form a complex. Further, we aimed to perform structural superimpositions automatically and to determine a threshold allowing the distinction between mutually exclusive or compatible interaction. Lastly, we aimed to find a good way of representing the results.

   Very importantly, all steps in this work should be accompanied by an extensive validation, improving previously described scoring systems and methods which thus would also contribute to increase the scientific community methodological knowledge.

CHAPTER

THREE

RESULTS

3.1 Structural and functional protein network analysis predicts novel signaling functions for the G-protein coupled receptor Rhodopsin

Rhodopsin, a G protein-coupled receptor (GPCR), is the major visual pigment in rod photoreceptor cells, responsible for the vision in dim light. Its activation by a photon of light triggers a signal transduction cascade which eventually leads to a hyperpolarization of the cell and then the transmission of the signal to the neural network of the retina. Rods are highly specialized cells. Phototransduction takes place in the rod outer segment (ROS) where a large number of disc membranes are tightly packed. These disc membranes contain many molecules of rhodopsin organized in dimeric conformation. These discs undertake an important turn over: they are being removed at the distal end by phagocytosis while new discs are generated at the base of the ROS. The core vision pathway is triggered by the activation of Rhodopsin by light. In this study, we were interested in exploring other pathways that could also be triggered by light-activated Rhodopsin.

We combined structural information with literature mining and proteomics data to construct the protein interaction network associated with rhodopsin. The proteomics data was of great value, since it provided all the components located in the highly specialized outer segment region of the rod photoreceptor and permitted

to restrict the literature mining and interaction information to this specific set of proteins. We further annotated the proteins according to their predicted functions, which allowed the decomposition of the network into functional sub-modules. We applied a structural analysis in a similar approach as described in section 3.2. This allowed us to annotate the connections among and between sub-modules using logical gates "AND" and "XOR", for compatible and mutually exclusive interactions respectively (Campagna *et al.*, 2008), involving 84 protein interaction structures from close homologs and 107 structurally predicted interactions using domain interaction structures significantly scored by InterPreTS. The annotation of these modules according to their physiological functions completed the process of building a high confident and comprehensive network with different levels of detail. We observed that "AND" gates are often occurring within stable molecular machines in the cell (typically involved in housekeeping, structure and polarity and metabolism pathways), whereas "XOR" gates are enriched within proteins involved in vesicle trafficking. The fact that proteins involved in vesicle trafficking can bind a limited number of protein simultaneously suggests a highly regulatory mechanism for protein transport and renewal of the ROS components. Interestingly, the vision pathway was connected to the other modules of the network by "XOR" gates, suggesting the local changes in concentration, resulting from the translocation of the vision protein between the outer and the inner segment (Reidel *et al.*, 2008), could be important for the transduction of the signal received by Rhodopsin.
Taken together, these results show that combining data sources are of great value to derive new insights from protein interactions network. Despite a low structural coverage in the Rhodopsin protein interaction network, we were able to gain insight about potential new routes activated by rhodopsin. This has been possible by the combination of data sources.

The article included in this section has been accepted for publication in *Molecular Systems Biology*.

3.1.1 Contributions

My contributions to this article are the following:

1. Comparison of the proteomics data sets.
2. Structural analysis of the protein interaction: structural modeling using 3DiD and InterPreTS, annotation of crystallographic artifacts, validations of the predictions.
3. Partial functional annotation of the proteins within the network.
4. Participation to the network reconstruction.

Structural and functional protein network analyses predict novel signaling functions for the G protein-coupled receptor rhodopsin

Christina Kiel[1#], Andreas Vogt[2,3#], Anne Campagna[1#], Andrew Chatr-aryamontri[4], Magdalena Swiatek-de Lange[2], Monika Beer[2], Sylvia Bolz[3], Andreas F. Mack[7], Norbert Kinkl[3], Gianni Cesareni[3,6*], Luis Serrano[1,5*], and Marius Ueffing[2,3*]

[1] CRG-EMBL System Biology Program, Centre de Regulació Genòmica (CRG) UPF, Dr. Aiguader 88, 08003 Barcelona, Spain

[2] Helmholtz Center Muenchen – German Research Center for Environmental Health, Department of Protein Science, 85764 Munich-Neuherberg, Germany

[3] University of Tuebingen, Division of Experimental Ophthalmology, Centre for Ophthalmology, Institute for Ophthalmic Research, Roentgenweg 11, 72076 Tuebingen, Germany

[4] University of Rome Tor Vergata, Department of Biology, 00133 Rome, Italy

[5] Institució Catalana de Recerca I Estudis Avançats (ICREA)

[6] Istituto Ricovero e Cura a Carattere Scientifico, Fondazione Santa Lucia, Rome

[7]University of Tuebingen, Institute of Anatomy, Am Oesterberg 3, 72074 Tuebingen

#These three authors contributed equally to this work

*Corresponding authors:

Marius Ueffing, Division of Experimental Ophthalmology, Centre for Ophthalmology, Institute for Ophthalmic Research, University of Tuebingen, Roentgenweg 11, D-72076 Tuebingen *and* Department of Protein Science, Helmholtz Center Muenchen, Ingolstaedter Landstr. 1, D-85764 Munich-Neuherberg, Germany. Tel: +49-7071 29-84021; E-mail: marius.ueffing@uni-tuebingen.de

Luis Serrano, CRG-EMBL System Biology Program, Centre de Regulació Genòmica, UPF. Dr. Aiguader 88, 08003 Barcelona, Spain. Tel.: +34-933-160-247; E-mail: luis.serrano@crg.eu

Abstract

Orchestration of signaling, photoreceptor structural integrity, and maintenance needed for mammalian vision remains enigmatic. By integrating three proteomic datasets, literature mining, computational analyses, and structural information, we have generated a multiscale signal transduction network linked to the visual G protein-coupled receptor (GPCR) rhodopsin, the major protein component of rod outer segments. This network was complemented by domain decomposition of protein-protein interactions and then qualified for mutually exclusive or mutually compatible interactions and ternary complex formation using structural data. The resulting information not only offers a comprehensive view of signal transduction induced by this GPCR but also suggests novel signaling routes to cytoskeleton dynamics and vesicular trafficking, predicting an important level of regulation through small GTPases. Further, it demonstrates a specific disease susceptibility of the core visual pathway due to the uniqueness of its components present mainly in the eye. As a comprehensive multiscale network, it can serve as a basis to elucidate the physiological principles of photoreceptor function, identify potential disease-associated genes and proteins, and guide the development of therapies that target specific branches of the signaling pathway.

Introduction

The work of many different groups over the past decades has led to a detailed understanding of the molecular mechanisms underlying the initial steps of the vision process in photoreceptor cells (Palczewski, 2006; Kwok *et al*, 2008, reviewed in Ridge *et al*, 2003). Rod photoreceptor cells are neurons capable of converting light into electrical signals. They possess a specialized structure consisting of five principal regions (Figure 1A): (i) the rod outer segment (ROS) composed of ~800 closed membrane discs where phototransduction takes place; (ii) the connecting cilium (CC) that joins the outer segment to the rest of the cell and regulates the traffic of proteins and other components in both directions; (iii) the rod inner segment (RIS), responsible for general cell metabolism, housekeeping, and protein production; (iv) the cell body with the nucleus (N); and (v) the synaptic region (SR) that makes the electrical connections to the neurons in the retina. Protein activity and turnover in the ROS are highly dynamic: about 10% of all discs are generated each day at the base of the segment, while older discs are removed at the distal end by phagocytosis of the neighboring retinal pigment epithelium cells (Boesze-Battaglia and Goldberg, 2002). To replenish, the components of the ROS and the vesicles synthesized in the RIS compartment need to be transported through the CC region, either actively or by diffusion (Reidel *et al*, 2008).

Rhodopsin is the major visual pigment in rod photoreceptor cells. It is a prototypical seven transmembrane-spanning G protein-coupled receptor (GPCR) that contains 11-cis-retinal as its intrinsic chromophore ligand, and it is highly concentrated in the ROS discs (Liang *et al*, 2003; Nickell *et al*, 2007). Due mainly to its high endogenous expression, rhodopsin was the first structurally resolved mammalian GPCR (Palczewski *et al,* 2000). In disc membranes, rhodopsin is tightly packed into paracrystalline dimer arrays, enabling optimal association with the heterotrimeric G-protein transducin as well as with additional regulatory components (Ciarkowski *et al*, 2005; Filipek *et al*, 2004; Fotiadis *et al*, 2004). Photon-activated rhodopsin promotes the activation of the associated G protein transducin, which in turn activates phosphodiesterase 6 (PDE6), leading to hydrolysis of cGMP and closure of the cGMP-gated channels. This initiates ultra-fast phototransduction (Hamer *et al*, 2005), translating light energy first into a biochemical signal, followed by an electrical cue that is transmitted through the neuronal network of

the retina. Adaptation to different light conditions, and regeneration of rhodopsin, is regulated at multiple levels, including through differential phosphorylation, differential calcium concentrations, and regulated enzymatic cycles, for example when regenerating 11-cis-retinal (Lamb and Pugh, 2004). Disruption of these highly organized structures and processes by germ line mutations can cause severe blinding diseases, such as retinitis pigmentosa, rod-cone dystrophies, and congenital stationary night blindness (Berger *et al*, 2010).

Proteomic analyses of purified ROS has identified about 500 proteins (Figure 1B) that include metabolic enzymes, transport proteins, cytoskeleton elements, regulatory proteins, scaffolds, and housekeeping components, providing a detailed description of the outer segment protein repertoire (Kwok *et al*, 2008) (Figure 1C). In addition, the relative abundance has been determined for 150 proteins (Kwok *et al*, 2008). Finally, studies have demonstrated that some of these proteins, such as arrestin, transducin, guanylate cyclase, and RhoA, localize differentially in the outer and inner segment in response to light/dark cycles (Hallet *et al*, 1996; Reidel *et al*, 2008; reviewed in Artemyev, 2008). To date, however, this wealth of data has not been fully analyzed nor integrated on a functional proteome-wide scale. It is anticipated that a large part of ROS proteins, and a core of the functional modules, will be common to all cells, while others will be photoreceptor-specific (Hofmann *et al*, 2006). Yet, similar to an assembly of music instruments, proteins organized as molecular machines can function in different, context-dependent ways. Connectivity, as well as the timing and tuning of different modules, appears to be crucial for the proper orchestration of signal transduction as well as for parallel signal processing in a concerted fashion: at the systems level, this results in the music of life. Although many details of the core phototransduction processes have been established, and mathematical models have been proposed (Dell'Orco *et al*, 2009), the overall orchestration of the outer segment functions, which include processes like disc shedding and renewal, protein transport along cilia, and light adaptation, are far from being understood. Further, how a variety of mutants of proteins primarily localized in the outer segments can cause visual impairment by perturbing the function of this organelle can only be speculated. For instance, a mutation could impair not only the proper folding of the protein but also its interactions with its partners within the physiologically functional

protein networks. Identifying protein interactions and their networks is therefore an important step toward improving our understanding of the molecular defects that underlie genetically-inherited and age-related blinding diseases, and may directly lead to identifying novel disease-associated genes.

There are several aspects that cannot be clearly determined through large-scale studies, such as the network dynamics, the simultaneous regulation of several distinct higher order biological outputs by one network, and the possibility that interactions detected for a particular protein might not be compatible simultaneously (Ho *et al*, 2002; Gavin *et al*, 2002; Gavin *et al*, 2006; Rual *et al*, 2005; Ito *et al*, 2001; Stelzl *et al*, 2005). As a consequence, information about the dynamics and the tempo-spatial resolution of networks has been limited to smaller signaling modules, such as receptor-initiated signal transduction (Olsen *et al*, 2006; Becker *et al*, 2010). Structural information can help discriminate between direct and indirect interactions in a given complex. More importantly, it can add a dynamical value to the classical interaction networks by determining if two or more predicted partners of any given protein or complex can simultaneously bind to a target, or if they instead compete for the same interaction surface (Kim *et al*, 2006; reviewed in Campagna *et al*, 2008 and Kiel *et al*, 2008). Integrating interaction data with protein expression information may assist in adding a dynamic dimension, and therefore a more realistic view, to the abstract "organism interactome" (Hofmann *et al*, 2006).

Here, we have combined experimental data, literature mining, and structural information to provide a comprehensive view of the signal transduction network centered on rhodopsin (see the flowchart in Figure 2). Integrating structural information with the relative estimates of expression levels allowed us to distinguish between mutually compatible or mutually exclusive interactions, enabling us to structure a network of nodes and edges towards sub-networks and functional modules. The resulting network offers an unprecedented view of signal transduction in vision and suggests a light-dependent orchestration of the core vision pathway to functions that have so far not been related to this pathway, such as cytoskeleton dynamics, vesicle transport, and energy metabolism. The specific light-dependent connectivity of rhodopsin to these functions is likely to be conferred by small GTPases and their regulators and interacting proteins,

such as the prenyl-binding protein PDEδ and other prenylated proteins. This would establish a dynamic and light-dependent mode of regulating the localization (to the membrane or cytosol), and thus the activity, of these GTPases.

Results

ROS proteome determination and contaminant removal

In order to determine the proteomic content of photoreceptor outer segments, dark-adapted porcine ROS and outer segments discs were isolated as previously described (Swiatek-de Lange *et al,* 2008; see Materials and methods). Proteins were then resolved by one-dimensional gel electrophoresis (1DE) and identified by mass spectrometry (MS). In three independent experiments, a total of 50 proteins were identified by MALDI-TOF from both ROS and ROS discs, and 434 proteins by the more sensitive Orbitrap-LC-MS/MS (Supplementary Table S1). The union of the two datasets resulted in a total of 444 proteins, of which 410 could be mapped to the human proteome. Comparing our dataset to a recent proteomic study that identified 516 proteins from bovine ROS (of which 487 mapped to the human proteome; Kwok *et al*, 2008) resulted in an overlap of 217 proteins (Figure 1B). We then created a unified dataset consisting of 680 human proteins, defined as our "*initial experimental ROS proteome"*.

To further refine the protein list presented in the *initial experimental ROS proteome*, we applied heuristic filtering procedures to remove proteins that might have contaminated the ROS fraction in either experiment (from the surrounding cells or from other cellular domains of the rod photoreceptor). First, we looked at the functional annotations of the *initial experimental ROS proteome* to identify annotations that contrasted with the expected properties of a ROS protein (such as transcription factors, nuclear proteins, and mitochondrial proteins), based on GO terms from the UniProt database (Supplementary Table S2). Second, we performed a detailed manual functional analysis based on UniProt, the KEGG database, and relevant literature. This revealed 81 putative contaminants (Supplementary Table S2); of these, 68 were found in only one of the two experimental datasets (e.g., ours and that from Kwok *et al*, 2008), further supporting their classification as contaminants. The 13 proteins identified in both sets are synaptic proteins and G proteins believed to be expressed only in cones (Kwok *et al*, 2008). We thus retained 605 proteins after these analyses.

We next removed all proteins for which there was no interaction data or further experimental evidence about their presence in ROS (protein group 1; Supplementary Table S2); this information was compiled from the literature and databases such as MINT

(for details, see Materials and methods and http://mint.bio.uniroma2.it/mint/) (Zanzoni *et al*, 2002; Chatr-aryamontri *et al*, 2007; Ceol *et al*, 2010; Supplementary Table S3). A total of 347 proteins passed this filter. From the analysis of published information, eight additional proteins that lacked associated protein-protein interactions were nevertheless considered to be *bona fide* ROS proteins with important functional and/or structural roles and were thus retained in the ROS proteome; these were the retinal-specific ATP-binding cassette transporter ABCA4, the cellular retinaldehyde-binding protein RLBP1, the photoreceptor outer segment all-trans-retinol dehydrogenase RDH8, peripherin-2 (PRPH2), the ROS membrane protein ROM1, Rab11B, retinitis pigmentosa 1 (RP1), and fascin 2 (FSCN2). This filtering procedure left us with 355 *bona fide* ROS proteins, or proteins that are dynamically localized to ROS (protein groups 2 and 3, respectively in Supplementary Table S2). These proteins represent the "*core ROS proteome"*.

Functional modules of the core ROS proteome

We next classified the *core ROS proteome* into six functional groups based on the above information, and we annotated lipid modifications, such as prenylation and geranylation (Supplementary Table S2 and Figure 1C): (1) *vision, signaling, transporters, and channels*: 56 proteins have functions that are either directly associated with vision or support visional functionality (i.e., visual cycle, protein homeostasis, or energy production). This module contains well-known members of the phototransduction pathway, including the core signal transduction of light (Dell'Orco *et al*, 2009) and the visual cycle involved in regenerating 11-cis retinal necessary to complement photosensitive rhodopsin after photo-bleaching (Lamb and Pugh, 2004). We further included here proteins involved in Ca^{2+}-dependent signaling and proteins associated with ion channels that regulate photoreceptor membrane conductance and polarity; (2) *outer segment structure and morphogenesis*: the seven proteins in this group are those implicated in outer segment structure and disc morphogenesis (Molday *et al*, 1987; Poetsch *et al*, 2001), and those that link the cytoskeleton to the extracellular matrix (ECM), such as alpha and beta catenin; (3) *housekeeping*: in this group of 73 proteins, we consider protein-folding chaperones and heat shock proteins, members of the ubiquitination/degradation-proteasome machineries, scaffold proteins such as the 14-3-3

family members, and proteins involved in oxidative stress, cell redox homeostasis, and apoptosis regulation (De La Paz and Anderson, 1992; for review, see Wenzel *et al*, 2005); (4) *cytoskeleton and polarity* (67 proteins): this group contains cytoskeleton proteins, such as actin and tubulin, as well as their respective binding proteins and molecular motors, proteins involved in regulating cytoskeleton dynamics including GTPases, and intermediate filaments. Many of these are associated with the connecting cilium and the axoneme of ROS and might therefore be present at low concentrations (reviewed in Adams *et al*, 2008). Proteins that are known to function in axon guidance were also added to this class; (5) *vesicles formation and trafficking*: we included here 60 proteins involved in Golgi function, protein and vesicle transport, and fusion, as well as the annexins that function in exocytosis and phagocytosis; (6) *metabolism*: we included here 92 proteins related to metabolism, in processes such as glycolysis, ATP synthesis, nucleotide, and fatty acid and carbon metabolism. Interestingly, we found that several of these are metabolic proteins involved in energy production (about 50% of the enzymes detected in this group are involved in glycolysis, and about 20% in the tricarboxylic acid pathway), including ATP synthase, the activity of which has recently been demonstrated in intact discs (Panfoli *et al*, 2009). This suggests that the ATP used in vision signaling is indeed produced within ROS and is probably fueled by glucose transported along the cilium. Indeed, this study identified a glucose transport protein, SLC2A1, as a ROS protein, supporting this hypothesis. High energy demands of ROS, and a capacity of only limited diffusion through the interconnecting cilium, may require on-site production of ATP.

Network reconstruction and structural modeling of the ROS interactome

Information about the protein-protein interactions (PPI) among the *core ROS proteome* was mined from protein interaction databases to assemble a ROS protein network (Supplementary Table S3); (Zanzoni *et al*, 2002; Chatr-aryamontri *et al*, 2007; Ceol *et al*, 2010; Kerrien *et al*, 2007). The protein interaction degree ranges from 1 to 179, with the highest number of interaction partners for actin, tubulin, 14-3-3 family members, heat shock protein members, and ERK (Supplementary Figure S1A).

Overall, the complete *core ROS proteome* PPI network consists of 5337 interactions among its members (Supplementary Table S3). The experimental evidence for most of these interactions (5047) was based on co-immunoprecipitation or pull-down experiments, which offers little support for their direct nature (Supplementary Figure S1B). In addition, many of the edges in the network are supported by single experimental pieces of evidence (> 85% of the PPI), often derived from high-throughput approaches. Thus, we refer to this network, which represents all the interactions that we could retrieve from published data, as a "*fuzzy ROS interactome*", since it contains many interactions supported by only one non-binary piece of evidence.

Next, we aimed at increasing the information content of the network by structural modeling. Pairs of interacting proteins often share common structural features with other interacting pairs of known structure (domains and linear motifs). We use structural information, combined with computational tools, to support low-confidence experimental interaction evidence and to determine whether two interactions involving a common partner are compatible or mutually exclusive (see Supplementary Material 1). We considered two levels of structural evidence that may support any given interaction. First, for each pair of members of the *core ROS proteome*, we searched the PDB database (http://www.pdb.org) for protein complexes of known structures whose elements share at least 70% homology with the query proteins. By this approach, we identified 84 complexes in the ROS core network whose structures could be confidently modeled on homologous structures (Supplementary Table S4). Most of the interactions for which there are X-ray structures, or structures from close homologs, are found between the connecting proteins in the modules 1 or 4 (*vision, signaling, transporters, and channels* and *cytoskeleton and polarity*) as well as among proteins involved in interactions connecting these two modules (Supplementary Figure S2).

Next, using a lower level of structural detail and confidence, we exploited the notion that similar domain pairs are likely to interact in a similar way ("nature repeats itself") (Aloy and Russell, 2002). For example, members of the Ras family and proteins containing a Ras-binding domain (RBD) are likely to use the same interaction surface when they interact (see Kiel and Serrano, 2006). To overlay a domain-level model on the ROS network, we represented each of the 355 nodes as a stack of Pfam domains

(http://pfam.janelia.org/) (Supplementary Table S5). We then searched the 3Did database (http://3did.irbbarcelona.org/; Stein *et al*, 2005; Stein *et al*, 2009) for structural evidence of pair-wise interactions between any of the domains in our database (for details, see Supplementary Material 1). Structural evidence was found for 352 pair-wise interactions, excluding pairs that had already been identified by comparison with homologous crystallized complexes (Supplementary Table S4). A confidence "interaction score" of ≥ 2.3 for the identification of the interacting pairs was obtained by interrogating the InterPReTS server (http://www.russelllab.or/cgi-bin/tools/interprets.pl/; Aloy and Russell, 2003). This score was validated using a yeast two-hybrid positive and negative binding data set described by Vidal and co-workers (Rual *et al*, 2005). We found a confidence of over 70% that two proteins containing the target domains will interact in a two-hybrid experiment when the InterPReTS score was above 2.3 (see Supplementary Material 1). Of the 352 interactions that had a hit in the 3Did database, 107 had InterPReTS scores higher than our chosen threshold and were therefore annotated as "supported by structural evidence". A total of 191 interactions supported by structural evidence (that is, the 84 interactions with known or closely related structures, and the 107 with significant InterPReTS scores) were merged with the literature-based interaction network. Interactions that could be annotated with structural evidence were mainly found within the functional modules (Supplementary Figure S2).

To increase the confidence in the resulting network, edges that were only supported by a single piece of evidence from any type of experiment except yeast two-hybrid experiments were removed (Supplementary Table S6), with the exception of interactions for which there was also structural information available (that is, a three-dimensional structure of the complex itself or of a highly homologous complex). This curated static network ("*high-confidence ROS interactome*") comprises 660 edges and links the majority of the nodes (with 266 proteins, as indicated in Supplementary Table S2; Figure 3A) that were present in the original network. The missing nodes are equally distributed among the proteins with respect to their GO terms, although an enrichment for proteins assigned to the classes *retinol recycling* and *metabolism* was observed (of 80% and 50%, respectively).

By considering only edges supported by at least one evidence of direct binary interaction, we obtained a "*high-confidence binary ROS interactome*" that contains 222 nodes (note that most of the nodes that were not captured by this network are annotated with metabolism ontology terms), linked by 349 edges (indicated as binary in Supplementary Table S6). Except for reactions involving guanylate kinase (GK) and nucleoside diphosphate kinase (NDPK), a nucleotide and cyclic nucleotide modifying enzyme, all interactions of the core vision pathway (Dell'Orco *et al*, 2009; Ridge *et al*, 2003; Wensel, 2008) are represented in our PPI network as true binary interactions. Thirty-five nodes have more than 10 interaction partners, with a maximum degree of 55. Only two proteins in the vision category have more than 10 interaction partners (CALM1 and CAMK2A). Most of the interactions involve the heat shock proteins, 14-3-3 family members, ERK (MAPK1), tubulins, and actin. More than 10 interaction partners were found in the metabolism branch for glyceraldehyde-3-phosphate dehydrogenase (GADPH) and for the two ATPase subunits.

Additionally, 109 direct binary interactions connected the defined functional modules, and 240 binary interactions connected proteins within modules (Figure 3B; for a detailed description of the connections between the modules, see the Supplementary Material 2). Out of those 240, 188 are classified within sub-boxes/sub-functions. The observation that roughly two-thirds of the interactions were found within functional modules, and only a third between modules, provides confidence to our module classification and manual functional annotation. Interestingly, the most highly connected modules are module 1 (vision) and module 4 (cytoskeleton), illustrating the important crosstalk between the core vision pathway and the cytoskeleton. The less connected modules are the ones involved in the structure of the discs and in metabolism. As expected, the housekeeping module, despite having fewer connections, is linked to all other modules.

The high-confidence ROS interactome suggests new functional links

Using our curated high-confidence binary ROS interactome as a basis, we decided to analyze in more depth the core vision pathway, which is probably one of the best-studied biochemical pathways (Ridge *et al*, 2003; Wensel, 2008; Dell'Orco *et al*, 2009). We

extended the published core vision pathway (Dell'Orco *et al*, 2009) using evidence from our high-confidence network and indicated structural coverage and outputs to different functional cellular processes emanating from the proteins in the pathway (Figure 4; for a detailed description, see Supplementary Material 2). Of these, we decided to validate the link to the GTPases RhoA and Rac1 (Figure 4 link A and Figure 4 link D), which suggests a link between vision activation and cytoskeleton reorganization.

1) Rho-Rac1 and the cytoskeleton connection

Previous work has demonstrated functional links between rhodopsin, certain GTPases (Mitchell *et al*, 1998) (most prominently transducin), and the cytoskeleton. S-arrestin specifically binds to activated and phosphorylated rhodopsin, inhibiting activation of transducin and terminating phototransduction (Kühn *et al*, 1978, Kühn *et al*, 1984, Wilden *et al*, 1986). Nair *et al*. have shown interactions between S-arrestin and microtubules (Figure 4, link G) (Nair *et al*, 2004).

We were able to confirm that small GTPases Rac and the GTP-bound form of RhoA bind rhodopsin, as has been previously described (Balasubramanian and Slepak, 2003; Wieland *et al*, 1990a; Wieland *et al*, 1990b; Gray *et al*, 2008) (Figure 4, link A). For this, we performed co-segregation/co-sedimentation experiments to reveal proteins within large complexes, as described previously to analyze the light-harvesting complex of photosystem II in plants (Swiatek-de Lange *et al*, 2008) (Supplementary Figures S3 and S4A). These experiments indicated that Rac1, Rho, and CRMP-2 were present in a large complex that also contained cytoskeletal proteins, rhodopsin, and components of the vision pathway. Although the low resolution of the technique, and the complexity of the patterns, preclude using these experiments to add new binary interactions to the ROS network, it can be used to corroborate interactions supported by further experimental evidence or from literature (Supplementary Tables S3 and S6). Using BN-PAGE (Schägger and von Jagow, 1991; Nijtmans *et al*, 2002; Camacho-Carvajal *et al*, 2004) or immunoprecipitation experiments in combination with either mass-spectrometry or subsequent immunoblotting, we obtained further evidence for the existence of large complexes containing rhodopsin that also included the cytoskeletal proteins actin and

tubulin as well as its specific regulators such a RhoA, Rac1, and CRMP-2 (Figure 5, Supplementary Figure S4B).

We next isolated the protein partners of the glycosylated N-terminus of rhodopsin by using concanavalin A affinity purification (De Grip, 1982; Plantner and Kean, 1976). The interactions between rhodopsin, RhoA, and CRMP-2 were confirmed by these concanavalin A pull-down experiments, and in part by additional co-immunoprecipitation experiments in which we detected rhodopsin associated with the core signaling complexes of the visual pathway including transducin and, again, with Rho, Rac1, and CRMP-2 (Supplementary Figure S4C and D). To confirm that CRMP-2, Rac1, and ROCK II are indeed bona fide ROS proteins rather than contaminants, we performed immunohistochemistry for these proteins. Indeed, all 3 proteins were constituents of ROS on cryosections of porcine retina (Figure 6). Despite considerable efforts, we were not able to confirm the presence of RhoA due to a lack of selectivity of various antibodies against RhoA in retinal sections.

2) *Functional analysis of the PDEδ-Rac1 complex*

PDEδ has been reported to bind to prenyl-modified proteins, such as several small GTPases and rhodopsin kinase (Hanzal-Bayer *et al*, 2002; Zhang *et al*, 2004), and it appears as an important node within our network (Figure 4, link D). PDEδ could thus play a critical regulatory role both in facilitating the transport of prenylated target proteins along the cilia together with Arl3 (Figure 4, link C) (Veltel and Wittinghofer, 2009) and in serving as an effector or guanine nucleotide dissociation inhibitor (GDI) for many GTPases, such as Arf, Rac1, RhoA, and Rab, all of which are expressed in photoreceptors (Figure 4, link D). Therefore, we tested whether PDEδ functions as a GDI for Rac1 in ROS. First, we demonstrated that PDEδ and Rac1 are colocalized in ROS using immunohistochemistry (Figure 6). Second, we showed that PDEδ and Rac1 colocalize in ROS in native protein complexes, by using dark-adapted ROS separated by BN-PAGE (Figure 7A). In dark-adapted ROS, PDEδ was part of distinct complexes that ranged from high molecular weight complexes of 660 kDa to smaller complexes of around 90 kDa, and interestingly, Rac1 colocalized with PDEδ within different complexes between the soluble and membranous fractions. Third, we tested whether

PDEδ could dissociate Rac1 from ROS membranes in vitro (see Materials and methods). Indeed, adding recombinant human (rh) PDEδ led to the solubilization of Rac1 from the ROS membranes (Figure 7B). Solubilization occurred in a dose-dependent manner with increasing amounts of rhPDEδ. As a positive control, we verified that PDEδ solubilized PDEβ from ROS membranes, as previously described (Florio *et al*, 1996). Thus, PDEδ can solubilize Rac1 from ROS membranes, a feature characteristic of GDIs.

All of the interactions determined here—with the exception of the ones identified by co-sedimentation, as this method is considered as weak evidence for physical interactions—were added as supporting evidence to our network (Supplementary Table S6). In total, co-purification and co-elution experiments supported 60 interactions that had been included in our network based on literature, and new evidence for 175 interactions from our co-immunoprecipitation results was added. Additionally, our results supported five interactions that had structural evidence (with INTERPRETS score ≥ 2.3). Restricted to a single new pathway (Rac1/RhoA–PDEδ–CRMP-2) our experimental data support the physiological relevance of our network. It should be noted, however, that this data cannot be considered to be complete or free of false positives, since the number of interactions tested and validated was small when considering the extent and complexity of the network.

Discussion

In this work, we investigated the protein interaction network in a highly specialized cellular region of the mammalian photoreceptors, the rod outer segment (ROS). Graphs representing protein interactions are idealized descriptions of all the interactions that can possibly occur in an organism. The realization that, in any given cell type, only a fraction of these interactions can possibly occur prompted the development of approaches to combine different genome-wide information to build interaction networks that are either specific for a cell type (Bossi and Lehner, 2009) or that change dynamically, such as during the cell cycle or after specific pathways have been induced. Here, we take this one step further and propose a protein interaction network for a structurally very distinct and functionally highly specialized region of the mammalian photoreceptors: the rod outer segment (ROS). In addition to proposing novel interactions, we present a structural model that allows us to discriminate between protein interactions that are compatible and those that are mutually exclusive.

A curated and structure-based PPI network central to rhodopsin

We first generated a ROS-specific protein interaction network by combining proteomic expression levels in ROS with interaction information, which we mined from the literature and then subsequently supplemented with our new data pertinent to the description of protein complexes in their physiological context. We next performed structural analysis of the curated network by decomposing proteins within this network into domains. This step allowed us to validate interactions at a domain level and to thereby increase the confidence in the network. By reassembling the decomposed network based on structural constraints into structure-functional modules, we were able to define logical relationships between the network nodes, and to define sub-networks that physically and functionally fit into molecular machines. Last, we annotated these functional modules according to their respective physiological processes, to derive a network of pathways and processes. Based on a compilation of experimental evidence and several layers of expert as well as automated curation, filtering, and modeling, the resulting network represents a multiscale description of wiring and physical connectivity

in the ROS of photoreceptors. The extended core pathway shows how rhodopsin activation-deactivation leads to other possible functional effects in addition to its primary function of signaling for closing the cGMP-gated cation channel. Thus, in addition to its relationship with the module of (1) *vision, signaling, transporters and channels*, wiring rhodopsin to (2) *outer segment structure and morphogenesis*, (3) *housekeeping*, and (4) *cytoskeleton and polarity* suggests a regulation of cytoskeleton assembly-disassembly and dynamics, vesicle and Golgi trafficking, and transport along the interconnecting cilium of photoreceptors by rhodopsin. Connections between active rhodopsin and Arf4 (Deretic *et al*, 2005; Mazelova *et al*, 2009), and between PDEδ and Rab13 and the GTP-bound form of Arl3 (Hanzal-Bayer *et al*, 2002), also link the vision cycle to vesicle trafficking and structure (Figure 4B and C). We experimentally validated two of the proposed new functional links. Our results suggest a link between rhodopsin, Rac1, RhoA, ROCK II, and CRMP-2. This points to a second, not yet experimentally tackled pathway that is influenced by light, which appears to be a delineation of an archetypical G-protein–regulated pathway known to be active in growth cone dynamics and collapse (Liu and Strittmatter, 2001). RhoA binds to CRMP-2 (gene name DPYSL2, Figure 5), a scaffold protein involved in actin cytoskeleton dynamics in neurons that regulates growth cone dynamics. CRMP-2, working through the GPCR lysophospatidic acid receptor, has been described as a crucial molecule in axon guidance, where it dynamically regulates the antagonistic effects of RhoA and Rac1. Regulated by a Rho-associated kinase (ROCK), CRMP-2 promotes either outgrowth or collapse in response to active RhoA or Rac1, respectively (Hall *et al*, 2001). When RhoA-GTP levels are high, more CRMP-2 is phosphorylated by the Rho-effector kinase ROCK, and thus less non-phosphorylated CRMP-2 is complexed with Rac1, leading to cytoskeleton collapse (reviewed in Liu and Strittmatter, 2001). CRMP-2 can bind directly to tubulin heterodimers to promote microtubule assembly (Fukata *et al*, 2002). This presents the exciting possibility that GPCR rhodopsin autoregulates its own axonal/dendritic guidance and possibly regulates outer segment growth via the archetypical mechanisms of axon guidance. Based on this scenario, the outer segment would function as a continuously extending growth cone, autoregulated by light and other as-yet unidentified guidance cues that may be produced in other retinal cells, most notably in the retinal pigment epithelium.

We additionally provide experimental evidence that PDEδ could act as a GDI for the small GTPase Rac1. PDEδ could thus play a very crucial regulatory role: (a) in transporting prenylated target proteins (Zhang *et al*, 2004) along the cilia, together with Arl3 (Veltel and Wittinghofer, 2009), and (b) as an effector or GDI for many GTPases (Hanzal-Bayer *et al*, 2002), such as Arf, Rac1, RhoA, and Rab, by keeping them GDP-bound and inactive. This is important since we did not find the conventional RhoGDI in ROS, suggesting that PDEδ could indeed substitute for this function in ROS (similar to that demonstrated for the small GTPase Rab13 in ROS; Marzesco *et al*, 1998). However, despite the structural similarity of the PDEδ and RhoGDI domains (Scheffzek *et al*, 2000; Hanzal-Bayer *et al*, 2002), we learned by superimposing the Rac1-RhoGDI with the Arl2-PDEδ structure that these two interactions depend on different moieties for binding (Supplementary Figure S6). We provide experimental evidence in this work that PDEδ could act as a GDI for Rac1. We did not find any GEFs or GAPs for small GTPases in our network but only GDIs (ARHGDI for RhoA, PDEδ for Rac1, and GDI1 and GDI2 for Rab proteins). Interestingly, this could suggest that these are not regulated by the usual switch-like mechanism of GTPase regulation, but rather by a gradient activation, in which the activity of active RhoA is determined only by the concentration of RhoGDI, keeping RhoA in the inactive form.

The role of Ca^2 in vision cycle, phototransduction, and actin cytoskeleton changes

Intracellular Ca^{2+} concentrations influence the activities of numerous kinases, such as different PKC isoforms, the PKA kinase, Ca^{2+}/calmodulin-dependent kinases, and the two CaMK-II isoforms, all of which are integral to the network. Predicted kinase phosphorylation sites from CaMK-II, PKA, PKC, MAPK, and PKD are summarized in Supplementary Table S7. Several Ca^{2+}-regulated kinases phoshorylate cytoskeletal target proteins, such as actinin and myosins, and small GTPases and their regulators. This opens the intriguing possibility that the nucleotide state and the dynamic spatial cellular distribution of several small GTPases are controlled by Ca^{2+}. As perturbed Ca^{2+} homeostasis is a consequence of the activity of a perturbed visual pathway in specific forms of retinitis pigmentosa (Paquet-Durand *et al*, 2010), this is likely to affect a variety of critical pathways and thus generate a systemic perturbance of ROS physiology. Our

network reveals several direct binary connections between Ca^{2+}-regulated proteins and cytoskeleton proteins: CaMK2A with actinin, calmodulin with GAP43 (neuromodulin) and S1008 (tubulin polymerization initiation), and PKC with 14-3-3 family members. Calmodulin is known to have a wide range of effector binding specificity, which dynamically changes with Ca^{2+} binding. Calmodulin 1 and 3 were linked to about 10 proteins from the two modules, cytoskeleton and vesicle transport. Calmodulin (CALM3 or CALM1) can bind to the cytoskeleton regulator spectrin alpha, actinin (ACTN2 and ACTN4), and the myosin motor protein MYO6. Therefore, calmodulin proteins could provide an important link between Ca^{2+}-signaling and regulation of the actin cytoskeleton, with spectrin playing a critical role in organizing and maintaining membrane sub-domains that harbor rhodopsin (Berghs *et al*, 2000). Further, a Ca^{2+}-dependent kinase, CaMK2A, was found to directly contact actinin-1, -2, and -4, and to be in a ternary complex with densin, a synaptic adhesion molecule (Walikonis *et al*, 2001), which is not present in our network (as it was not taken into consideration). Another link appears between calmodulin and RalA and RalB, both of which are involved in trafficking: RalA plays a role in exocytosis regulating exocyst assembly, while RalB interacts with EXOC8 (a part of the exocyst complex); RALBP1 is an effector of both RalA and RalB. Ca^{2+} activity is also likely to regulate metabolic activities through IHD3A, recoverin, and neurocalcin: the hippocalcin-like protein 1 is a recoverin-like protein that was suggested to have an anti-apoptotic function and might protect photoreceptors from Ca^{2+}-induced cell death (Krishnan *et al*, 2009).

Finally, Ca^{2+} could play an important role in the light-dark cycle by affecting PKA activity. Phosphorylation of RGS9-1 by PKA (Balasubramanian *et al*, 2001) is regulated by light and Ca^{2+}, and results in the reduction of RGS9-1 GAP activity: with light, RGS9-1 causes rapid Tα-GTP inactivation and photoreceptor recovery, while in the dark, PKA is activated by rising concentrations of Ca^{2+} and cAMP, which in turn phosphorylates RGS9-1. In this way, GAP activity is reduced, the active transducin lifetime is prolonged, and the photoresponse is strengthened (Balasubramanian *et al*, 2001).

While it remains to be seen how all of these connections are orchestrated, and to which degree they impact vision homeostasis, there is no doubt that Ca^{2+} plays a crucial role in ROS functionality.

Structural information, structural coverage, and "AND" and "XOR" gates

Structural information allows the confidence of any independent interaction evidence to be tested and at the same time can add topological information to the molecular level by defining sites or interaction domains. When several proteins can bind to a single protein, the various interactions can occur simultaneously or can be mutually exclusive (reviewed in Santonico *et al*, 2005; Kim *et al*, 2006). If two or more proteins compete for the same binding site, it seems unlikely that binding can occur simultaneously, whereas binding to topologically distinct sites may occur at the same time. At the level of graph representations within a network, structural information can thus support logical constraints. Here, it is important to mention that interactions were defined as exclusive or compatible from a structural point of view, and that this cannot be directly translated to biological terms in all cases (i.e. for competition to occur, the target protein should be at lower concentration than the competing ones). When both competitors are present at the same place and time, changes in concentration levels or additional regulatory constraints (for example, those introduced by post-translational modifications) could regulate competition.

Structurally superimposing domains onto interactions allowed us to define ternary complex formation and, importantly, to model both the composition of macromolecular assemblies and its dynamic dissection into mutually exclusive complexes (Supplementary Figure S7). With this information, we can add dynamics to the network, using the following "AND" and "XOR" ("XOR" = exclusive OR) logical gate symbols: if three or more proteins can interact at the same time, they are compatible (indicated with "AND"), while if three or more proteins cannot interact simultaneously, they are mutually exclusive (indicated with "XOR") (Figure 8). Competitors are frequently found in highly dynamic processes or may dynamically connect a given protein to different signaling and functional modules. The structural and interaction analyses of the core vision pathway and its cytoskeleton branch show several examples of non-compatible ("XOR") interactions (Figure 8). For example, rhodopsin may interact with transducin, arrestin, or rhodopsin kinase (in the core vision pathway). It may also interact with Rac1 or RhoA (which are antagonists in cytoskeletal dynamics) or with Arf4 (involved in trafficking).

Changes in rhodopsin activation, concentration, and localization, or in its activation states, may therefore switch signaling into different pathways. Further, rhodopsin localization during ciliary transport and disk formation, and dynamic changes in concentrations and activation states in response to light, can alter the array of rhodopsin binding partners, since these are determined by the phosphorylation state of rhodopsin on the one hand and the availability or concentration of binding proteins on the other hand. Interestingly, "AND" gates are mainly found in the *housekeeping*, *structure and polarity*, and *metabolism* branches, e.g. within large functional complexes, such as the T-complex, the proteasome, tubulin, and the ATP synthase machinery. "XOR" gates, which are mainly prevalent in the *vesicle structure and trafficking* branch, indicate switch behavior or redundant protein functions, such as for Rab GTPases (Del Conte-Zerial *et al*, 2008). In the *vision* branch, both "AND" and "XOR" gates synergize. This may allow dynamic tuning of light and dark states. However, all connections from the *vision* module to other modules are "XOR" connections, suggesting that competition, together with local protein concentration changes, could be important for transmitting signals from the core *vision* module.

The vision network and disease

A large fraction of retinopathies involve the degeneration of rod photoreceptors; these include retinitis pigmentosa (RP), syndromes incorporating retinal degeneration with different associated phenotypes (such as Usher syndrome and Bardet–Biedl syndrome), and Leber congenital amaurosis (LCA), a congenital form of retinal degeneration. An increasing number of genes and proteins has been implicated in these pathologies (http://www.sph.uth.tmc.edu/Retnet/). These proteins include: components of the visual transduction cycle; structural components of the cytoskeleton, rod and/or cone photoreceptor outer segment disc membranes; components of synthesis and recycling of the retinoid; transcription factors (including CRX and NRL) and splicing factors; those involved in signaling and cilium maintenance, phagocytosis of the outer segment discs of the photoreceptors, and trafficking of intracellular proteins; and those with functions in pH maintenance in the retina, in metabolism, and as chaperones. The protein with by far the largest number of mutations is rhodopsin (>100 mutants), while the others range

contain from 40 mutations (for the retina-specific crumbs homolog 1 [CRB1]) to 1 (for transducin alpha) (see Supplementary Material 3). Structural analyses of the different mutations mapped on the available structures or homology models (156 mutations) indicated that the majority of these are within the hydrophobic core of the corresponding proteins and are therefore likely to cause misfolding (see Supplementary Material 3). Mapping all proteins involved in vision-related diseases into the network made it apparent that the core visual pathway is the most susceptible to disease, and that the other functional modules are relatively robust. Out of 36 proteins considered here to be involved in retinal degeneration, the majority (20 proteins) are localized in ROS (the remaining are found in other regions of the rod cells or in other cells involved in retina homeostasis [pH control], retinal recycling, or phagocytosis of the ROS discs). We found two cases of a ROS protein also expressed in other tissues, with no other apparent phenotype; for example, isocitrate dehydrogenase NAD-dependent subunit B is found in many cell types besides rod cells.

The prevalence of proteins from the core visual pathway in disease may have several explanations: first, mutations in other modules central for cellular function may result in a systemic all-or-nothing behavior, affecting the overall viability or proper development of an organism and thereby causing early death. This may be true for critical cytoskeletal proteins and GTPases, as for example those involved in vesicle trafficking and maturation, and for proteins involved in metabolic activity. Second, the lack of redundancy for the very specific functions within the visual pathway might cause it to be more susceptible. Here, evolution may have favored high-end functional properties over the robustness of the pathway. Thus, lack of redundancy may have been accepted by evolution even though it interferes with robustness as a pay-off for the high-end performance that is achieved in photoreceptors with single photon detection and with multi-color vision.

Conclusions

Taken together, this work suggests that rhodopsin is able to trigger several distinct physiological activities in addition to its primary function of closing and opening the camp-gated cation channel. Considering protein interactions as a result of domain

interactions has allowed us to increase the resolution, define discrete functional modules, and add a spatial dimension to this network. Based on this study, we obtained a novel biological insight that offers new testable hypotheses, which have been partially validated through the experiments performed here, namely, of the connectivity of rhodopsin to small GTPases involved in cytoskeleton assembly/disassembly and dynamics, and to vesicle and Golgi trafficking. This suggests a role for rhodopsin in self-regulating and fine-tuning the structural and functional integrity of photoreceptors. Cytoskeleton changes, such as microtubule assembly reorganization, are likely to affect protein transport between the inner and outer segments during light-to-dark changes (reviewed in Reidel *et al*, 2008), as well as to regulate cell polarity and disc development. The involvement of rhodopsin in regulating intracellular Ca^{2+} levels suggests its role in an overarching Ca^{2+}-dependent regulatory network that determines dynamic changes in kinase activity and protein complex assembly. This in turn results in higher-order physiological behavior, such as cytoskeletal dynamics and vesicular trafficking tuned by light. At a systems level, these network relationships imply a concerted regulation of outer segment structure, polarity, and vesicular trafficking orchestrated via GTPase-guided signaling pathways activated by light, Ca^{2+}-regulated processes activated by cGMP gated channel activity (and thus also by light), and cytoskeletal and ciliary dynamics (which may also be fine-tuned by light). With respect to disease, we can conclude that, among at least four pathways driven or regulated by rhodopsin, the visual pathway is the only one highly associated with disease, whereas all others are relatively unaffected. Conceptually, our work presents a general approach applicable to the analysis of any cellular pathway. The resulting comprehensive multiscale “vision network” can serve as a basis for elucidating physiological principles of photoreceptor function and may help to identify potential disease-associated proteins and to guide signaling branch-specific therapy development.

Materials and methods

Isolation of ROS and ROS discs

Porcine eyes were obtained from a local slaughterhouse. After the retinae were dissected, two approaches for ROS isolation were compared: that to Molday (Molday *et al*, 1987) with that of Papermaster (Papermaster and Dreyer, 1974). Briefly, for the Molday protocol, ROS were detached from the retinal tissue by gentle mechanical homogenization in cold isolation medium (20% [w/v] sucrose, 20 mM Tris, 2 mM $MgCl_2$, 130 mM NaCl, at pH 7.2) and separated from the homogenate by loading onto a 27-50% linear sucrose density gradient. Alternatively, fresh retinae were homogenized by shaking in cold isolation medium (34% [w/v] sucrose, 65 mM NaCl, 2 mM $MgCl_2$, and 5 mM Tris-acetate buffer, pH 7.4). ROS were then pelleted by centrifugation, and the remaining retinal tissue was re-homogenized with a teflon homogenizer. Supernatants from both homogenization steps (crude ROS) were combined and loaded onto step-density gradients of 1.15, 1.13, and 1.11 g/ml sucrose. After cold centrifugation in a Beckman SW40-rotor for 1 h at 38 000 rpm, purified ROS were collected from the surface of a 1.11-1.13 g/ml sucrose gradient, and the protein content was determined by Bradford assay (BioRad). Osmotically intact discs were isolated from ROS according to Smith (Smith *et al*, 1975). ROS were ruptured by osmotic shock and intact discs were separated by flotation in 10% Ficoll (Sigma). After centrifugation (120 000 × g, 2 h, 4°C) intact discs were harvested from the Ficoll surface. The purity of the ROS preparations was either checked optically by microscope (Figure 1A inset) or by immunoblot analysis for RIS markers (BIP and Tom20) (Supplementary Figure S8).

Sucrose density gradient centrifugation

ROS or intact discs were ruptured by osmotic shock, and the membranes were separated from soluble fraction by centrifugation. An amount of membrane equivalent to 1 mg protein was solubilized in 1% (w/v) β-dodecylmaltoside (Sigma) as described (Mueller and Eichacker, 1999), loaded onto linear 0.1-1.0 M sucrose gradients, and centrifuged for 17 h at 230 000 × g at 4°C. Individual gradient fractions were either loaded directly for

SDS-PAGE or were precipitated with methanol/chloroform as previously described (Wessel and Flügge, 1984).

SDS-PAGE and immunoblotting

SDS-PAGE and subsequent immunoblotting on PVDF membranes (Amersham) were carried according to standard procedures. Antibody-antigen complexes were visualized using enhanced chemiluminescence detection (ECL+, Amersham) on Hyperfilm (Amersham). Immunoblots were incubated with the following antibodies: anti-RhoA 26C4 and anti-ROCK II H-85 (Santa Cruz), anti-visual arrestin, anti-transducin alpha, and anti-rhodopsin (Affinity BioReagents), anti-rhodopsin (Acris Antibodies), anti-CRMP-2 (C4G, a generous gift from M. Morishima and Y. Ihara, University of Tokyo, Japan), anti-Rac1 (BD Transduction Laboratories), anti-RhoABC (Sigma), anti-BIP (BD Bioscience), and anti-Tom20 (BD Bioscience). HRP-coupled secondary goat-anti-rabbit and goat-anti-mouse antibodies were obtained from Jackson ImmunoResearch.

Immunoprecipitation

Immunoprecipitation (IP) was performed with anti-RhoA-agarose- or anti-Rac1-agarose-conjugated antibodies (Santa Cruz) or anti-rhodopsin (Acris Antibodies). An amount of ROS equivalent to 500 µg protein was ruptured by osmotic shock in lysis buffer (50 mM NaCl, 1 mM EDTA, 20 mM Tris-HCl, pH 6.8) and centrifuged to separate the membrane and soluble fractions. The membrane fraction was solubilized in 1% (w/v) β-dodecylmaltoside (DM), and the soluble fraction was directly subjected to IP. For anti-rhodopsin IP, solubilized ROS (1% DM) was directly subjected to IP. Nonspecific protein binding to agarose beads was prevented by pre-incubation of the fraction with 25% protein G-agarose (Santa Cruz). For IP, the ROS fractions were incubated with 5-10 µg of antibody conjugate/antibody at 4°C for 3 h or overnight with rotation. As a control for nonspecific antibody binding species, specific IgGs (Sigma Aldrich) were used.

Immunohistochemistry

Porcine eyes were obtained from a local slaughterhouse, fixed in 4% paraformaldehyde in 0.1 M phosphate buffer (PB) for 4 h, and rinsed in 0.1 M phosphate-buffered saline

(PBS). Cornea, lens, and vitreous body were removed, and the retina was cut in 1.5 × 1.5 cm pieces. The fixed tissue was cryoprotected at 4°C step-wise in 10%, 20%, and 30% sucrose in PBS, for 1 h for the first two steps and overnight for the last step. Retina was then embedded in tissue-freezing medium (LeicaMicrosystems) and frozen in liquid nitrogen. 12 µm sections were prepared, mounted on Superfrost glass slides, and air dried at 37°C. Retinal sections were rinsed in PBS and then non-specific binding sites were blocked with PBS containing 10% normal goat serum (NGS), 1% bovine serum albumin (BSA), and 0.3% Triton X-100 for 1 h at room temperature. Sections were incubated overnight at 4°C with the following primary antibodies diluted in blocking solution: rabbit anti-CRMP-2 (1:300; Abcam), rabbit anti-Rac1 (1:100; Sigma), rabbit anti-ROCK II (1:300; Abcam), rabbit anti-PDE6δ (1:200; ABR), or mouse anti-rhodopsin (1:200; Millipore). Sections were then washed in PBS and incubated with the appropriate fluorescent-labeled secondary antibody (goat anti-rabbit IgG-Alexa 568 or goat anti-mouse IgG-Alexa 568; Molecular Probes) diluted 1:500 in PBS for 1 h at room temperature. Nuclei were counterstained with Sytox Green Nucleic Acid Stain (Molecular Probes). After three final washes in PBS, sections were mounted with Mowiol 4-88 (Polysciences). As negative controls, the primary antibodies were also omitted; in these cases, no staining was observed. Stained cryostat sections were analyzed and scanned with a confocal laser scanning microscope (Zeiss LSM510 META, Jena, Germany), using an argon laser at 488 nm and a He/Ne laser at 543 nm excitation with appropriate filter sets. Images were taken sequentially to assure that only one channel was detected at a time. The Sytox Green nuclear stain was allocated to the blue color channel for convenient viewing. Transmitted light images with DIC optics (Nomarski) were recorded simultaneously. Control sections without primary antibody incubation were scanned with the same laser and detection settings.

Concanavalin A pull-down

For concanavalin A pull-down experiments, an amount of ROS equivalent to 300 µg protein was ruptured by osmotic shock in lysis buffer. The membrane ROS fraction was solubilized in 1% (w/v) β-dodecylmaltoside (Sigma Aldrich), and the soluble fraction was directly subjected to concanavalin A pull-down. The ROS fractions were incubated

with 50 µl of concanavalin A sepharose (Amersham Biosciences) conjugate for 3 h at 4°C. Nonspecific protein binding of the rhodopsin-associated protein to concanavalin A was prevented by performing the pull-down in the presence of 0.2 mM α-methylmannoside (as the presence of α-methylmannoside lowers the affinity of proteins for the beads).

Blue-native PAGE

Membranes from either ROS or intact discs corresponding to 300 µg protein were suspended in 60 µl buffer containing 750 mM ε-aminocapronic acid, 50 mM bis-Tris, pH 7.0, and 0.5 mM EDTA, and then solubilized in 1% (w/v) β-dodecylmaltoside. The solubilized membrane samples were added to a buffer containing 5% (w/v) Serva Blue G in 750 mM ε-aminocapronic acid, loaded onto 4-12% PAA gradient gels, and electrophoresed (Schägger and von Jagow, 1991). To separate in the second dimension, gel lanes were incubated for 20 min in solubilization buffer containing 2% (w/v) SDS, 66 mM DTT, and 66 mM Na_2CO_3, and loaded onto denaturing PAA gels.

MS–MALDI-TOF

Selected spots were excised from dried silver-stained gels, destained (Gharahdaghi *et al*, 1999), dehydrated in 40% acetonitrile (100 µl), and subjected to tryptic proteolysis in 1 mM Tris-HCl, pH 7.5, and 0.01 µg/µl trypsin. In parallel studies, proteins excised from dried gels were subjected to SDS removal by ion-pair extraction prior to in-gel tryptic proteolysis as described (Zischka *et al*, 2004). MALDI-TOF PMFs were obtained on a Bruker Reflex III mass spectrometer (Bruker Daltonics, Bremen). Aliquots from each tryptic digest were co-crystallized with a matrix composed of 2.5-dihydroxybenzoic acid (20 mg/ml in 20% acetonitrile, 0.1% trifluoroacetic acid [TFA]) and 2-hydroxy-5-methoxybenzoic acid (20 mg/ml in 20% acetonitrile, 0.1% TFA) in a 9:1 ratio (v/v) on 400 µm AnchorChipTM targets (Bruker Daltonics). Alternatively, PMF and MS/MS spectra were measured on AB4700 mass spectrometer (Applied Biosystems, Darmstadt, Germany), and aliquots from each tryptic digest were co-crystallized with a matrix comprised of 5% cyanohydroxycinnamic acid (in 70% acetonitrile, 0.1% TFA) on steel

targets (Applied Biosystems). Database searches were performed using the Mascot software (Perkins *et al*, 1999).

MS–Orbitrap

LC-MS/MS analysis was performed on an Ultimate3000 nano-HPLC system (Dionex) coupled to a LTQ OrbitrapXL mass spectrometer (Thermo Fisher Scientific) by a nanospray ion source. Tryptic peptide mixtures were automatically injected and loaded with a flow rate of 30 µl/min in 95% buffer C (0.5% TFA in HPLC-grade water) and 5% buffer B (98% actetonitrile, 0.1% formic acid in HPLC-grade water) onto a nanotrap column (100 µm i.d. × 2 cm, packed with Acclaim PepMap100 C18, 5 µm, 100 Å, Dionex). After 5 min, peptides were eluted and separated on an analytical column (75 µm i.d. × 15 cm, Acclaim PepMap100 C18, 3 µm, 100 Å, Dionex) by a linear gradient from 5% to 40% of buffer B in buffer A (2% acetonitrile, 0.1% formic acid in HPLC-grade water) at a flow rate of 300 nl/min over 90 min. The remaining peptides were eluted by a short gradient of 40% to 100% buffer B over 5 min. Eluting peptides were analyzed by the LTQ OrbitrapXL. From the high resolution MS pre-scan with a mass range of 300 to 1500, the ten most intense peptide ions were selected for fragment analysis in the linear ion trap if they exceeded an intensity of at least 200 counts and if they were at least doubly charged. The normalized collision energy for CID was set to a value of 35, and the resulting fragments were detected with normal resolution in the linear ion trap. The lock mass option was activated, and a background signal of a mass of 445.12002 was used for the lock mass. Every ion selected for fragmentation was excluded for 30 seconds by dynamic exclusion. The raw data was analyzed using Sequest (Thermo Fisher Scientific) and Scaffold (Proteome Software) as described previously (Gloeckner *et al*, 2009) against a non-redundant pig, human, mouse, rat, and bovine protein sequence database derived in-house from Uniref100, due to a insufficient number of entries for porcine proteins in the databases. Proteins were considered to be specific when they displayed two or more peptides (with a peptide probability >95%) in at least two out of four experiments. The protein probability threshold was set to 99%. Contaminants such as keratins were removed.

Comparison of different proteomic data sets determined in ROS

All proteins identified in the three different proteomic datasets were mapped to their corresponding human ortholog gene IDs by sequence comparison (using the default BLAST value of 10) and then compared. Since the proteomic analysis of Liu *et al* (Liu *et al*, 2007) also contains part of the cilium, our "near-to-complete" proteomic data set was defined as the union of the protein set identified by Kwok *et al* (2008) with the one determined here (Figure 1B).

Protein interaction network analysis

All the results described in our studies were uploaded into Supplementary Tables S3 and S6 according to standard database curation rules (Ceol *et al*, 2010; Zanzoni *et al*, 2002). Results from pull-down and co-immunoprecipitation experiments were resolved as binary protein interactions, in which each bait protein was linked to all identified preys. Co-sedimentation and complex-purification experiments that unambiguously identified protein complexes but lacked sufficient detail to determine their exact interaction topology are represented in the database as a list of interactors (complex members). A comprehensive literature mining and database curation effort was also carried out in order to include as exhaustively as possible the set of rhodopsin/vision related interactions already described in the scientific literature. The curated interaction sets were represented and analyzed by the Cytoscape visualization and analysis software (Shannon *et al*, 2003). PPI data from databases included interactions determined from ROS extracts (by co-sedimentaion or affinity chromatography). However, in the majority of cases, interaction information was derived from *in vitro* experiments, such as large-scale yeast two-hybrid screens, or tandem-affinity purifications in artificial cell systems. Additional, low-scale PPI data from literature include data determined with quantitative affinity methods, such as isothermal titration calorimetry, surface plasmon resonance, nuclear magnetic resonance, and peptide arrays (using purified proteins), or PPI data from non-quantitative methods, such as affinity chromatography (GST pull-down), crosslinking, and enzyme assays. According to the MINT curation rules, interactions were considered to be direct if they were supported with evidence obtained with one of the following methods, as described in the PSI MI controlled vocabulary: two-hybrid, enzymatic studies, two-

hybrid pooling approach, two-hybrid array, beta lactamase complementation, surface plasmon resonance, fluorescence resonance energy transfer, biochemical, biophysical, protein arrays, protease assays, bimolecular fluorescence complementation, far-western blotting, cross-linking studies, electron paramagnetic resonance, two-hybrid fragment pooling approach, protein kinase assay, GTPase assay, enzyme-linked immunosorbent assays, peptide arrays, isothermal titration calorimetry, bioluminescence resonance energy transfer, competition binding, fluorescence technologies, antibody arrays, saturation binding, fluorescence polarization spectroscopy, protease accessibility laddering, affinity technologies, protein cross-linking with a bifunctional reagent, ubiquitin reconstruction, fluorescence microscopy, beta galactosidase complementation, biochemical activity, classical fluorescence spectroscopy, fluorescence technology, phosphatase assay, and reconstituted complex.

Structural information and interaction modelling

Structural information was derived by a combined approach of comparing different domain interaction types, as listed in the 3Did database (http://3did.irbbarcelona.org/; Stein *et al*, 2005; Stein *et al*, 2009), between two interacting proteins for which there was experimental evidence that they could form a complex. The 3DID database was improved by analyzing all structures for crystallographic artifacts, using: (i) the interaction annotation of the author, or, if this was not available (ii) the protein quarternary structure (PQS) method (Henrick and Thornton, 1998). The confidence of two domains to mediate the interaction was then assessed using the InterPreTS (http://www.russelllab.or/cgi-bin/tools/interprets.pl/; Aloy and Russell, 2003) scoring system, which evaluated sequence similarity and amino acid propensities in the interface. We further screened for all X-ray and NMR complex structures and homologs (with a sequence similarity threshold of 70%) among all 360 proteins of the network. For further details, see the Supplementary Material 1.

PDEδ subunit activity assay

Recombinant PDEδ protein (rhPDEδ) was obtained from GenWay Biotech at a concentration of 0.7 μg/μL in storage buffer (10 mM Tris, pH 8.0, 0.1% Triton X-100,

0.002% NaN_3, and 10 mM dithiothreitol). An amount of ROS (in isolation medium) corresponding to 100 µg protein was ruptured by three freeze-thaw cycles in liquid nitrogen and centrifuged at 4°C for 30 min at 100 000 g (Beckman Optima ultracentrifuge; Rotor TLA110). The resulting pellet, containing the membranous fraction, was resuspended in 100 µL incubation buffer (25 mM Hepes, 20 mM Tris-HCl, pH 7.5, 1 mM dithiothreitol, 1 mM $MgCl_2$, 5 mM EDTA, 150 mM NaCl, and protease inhibitor cocktail [Roche]) and then incubated with different amounts (0, 0.5, 1, 2, 4, 6 or 8 µg) of rhPDEδ for 1 h at 37°C in a horizontal shaker. To rule out that the Triton X-100 in the storage media affected the recombinant PDEδ, all samples were adjusted to the same volume (volume of the sample with the highest PDEδ concentration used) with storage buffer. Samples were separated into membrane and soluble fractions by centrifugation at 4°C for 30 min at 100 000 g and analyzed by SDS-PAGE and Western blot using anti-Rac1 antibodies and anti-PDEδ antibodies.

Acknowledgements

This work was supported by the German Federal Ministry for Education and Research through the BMBF grant QuantPro 0316865A, Dynamo 0315513A, and BMBF – IMAGING FKZ 0315508A, to MU. We like to acknowledge the EU funding for financing part of the work: INTERACTION PROTEOME, LSHG-CT-2003-505520 (to MU, GC and LS), PROSPECTS, grant agreement number HEALTH-F4-2008-201648 (to LS) and SYSCILIA, grant agreement number HEALTH-F5-2010 -241955 (to MU). MU and GC received funding from the European Community's Seventh Framework Programme FP7/2009 under grant agreement number 241481, AFFINOMICS.

Author contributions

CK, LS, GC, and MU designed the research. CK, LS, GC, AV and MU wrote the paper. AV, MS, MB, SB, AM performed experiments. CK, ACa, AV, NK, ACh, GC, LS, and MU analyzed the data and/or provided data analyses expertise.

References

Adams M, Smith UM, Logan CV, Johnson CA (2008) Recent advances in the molecular pathology, cell biology and genetics of ciliopathies. *J Med Genet* **45**: 257-267

Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A* **99**: 5896-5901

Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* **19**: 161-162

Artemyev NO (2008) Light-dependent compartmentalization of transducin in rod photoreceptors. *Mol Neurobiol* **37**: 44-51

Balasubramanian N, Levay K, Keren-Raifman T, Faurobert E, Slepak VZ (2001) Phosphorylation of the regulator of G protein signaling RGS9-1 by protein kinase A is a potential mechanism of light- and Ca2+-mediated regulation of G protein function in photoreceptors. *Biochemistry* **40**: 12619-12627

Balasubramanian N, Slepak VZ (2003) Light-mediated activation of Rac-1 in photoreceptor outer segments. *Curr Biol* **13**: 1306-1310

Becker V, Schilling M, Bachmann J, Baumann U, Raue A, Maiwald T, Timmer J, Klingmüller U (2010) Covering a broad dynamic range: information processing at the erythropoietin receptor. *Science* **328**: 1404-1408

Berger W, Kloeckener-Gruissem B, Neidhardt J (2010) The molecular basis of human retinal and vitreoretinal diseases. *Prog Retin Eye Res* **29**: 335-375

Berghs S, Aggujaro D, Dirkx R Jr, Maksimova E, Stabach P, Hermel JM, Zhang JP, Philbrick W, Slepnev V, Ort T, Solimena M (2000) BetaIV spectrin, a new spectrin localized at axon initial segments and nodes of ranvier in the central and peripheral nervous system. *J Cell Biol* **151**: 985-1002

Boesze-Battaglia K, Goldberg AF (2002) Photoreceptor renewal: a role for peripherin/rds. *Int Rev Cytol* **217**: 183-225

Bossi A, Lehner B (2009) Tissue specificity and the human protein interaction network. *Mol Syst Biol* **5**: 260

Camacho-Carvajal MM, Wollscheid B, Aebersold R, Steimle V, Schamel WW (2004) Two-dimensional Blue native/SDS gel electrophoresis of multi-protein complexes from whole cellular lysates: a proteomics approach. *Mol Cell Proteomics* **3**: 176-182

Campagna A, Serrano L, Kiel C (2008) Shaping dots and lines: adding modularity into protein interaction networks using structural information. *FEBS Lett* **582**: 1231-1236

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res* **35**: D572-D574

Ciarkowski J, Witt M, Slusarz R (2005) A hypothesis for GPCR activation. *J Mol Model* **11**: 407-415

Ceol A, Chatr-aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* **38**: D532-D539

De Grip WJ (1982) Purification of bovine rhodopsin over concanavalin A-sepharose. *Methods Enzymol* **81**: 197-207

De La Paz MA, Anderson RE (1992) Lipid peroxidation in rod outer segments. Role of hydroxyl radical and lipid hydroperoxides. *Invest Ophthalmol Vis Sci* **33**: 2091-2096

Del Conte-Zerial P, Brusch L, Rink JC, Collinet C, Kalaidzidis Y, Zerial M, Deutsch A (2008) Membrane identity and GTPase cascades regulated by toggle and cut-out switches. *Mol Syst Biol* **4**: 206

Dell'Orco D, Schmidt H, Mariani S, Fanelli F (2009) Network-level analysis of light adaptation in rod cells under normal and altered conditions. *Mol Biosyst* **5**: 1232-1246

Deretic D, Williams AH, Ransom N, Morel V, Hargrave PA, Arendt A (2005) Rhodopsin C terminus, the site of mutations causing retinal disease, regulates trafficking by binding to ADP-ribosylation factor 4 (ARF4). *Proc Natl Acad Sci USA* **102**:3301-3306

Filipek S, Krzysko KA, Fotiadis D, Liang Y, Saperstein DA, Engel A, Palczewski K (2004) A concept for G protein activation by G protein-coupled receptor dimers: the transducin/rhodopsin interface. *Photochem Photobiol Sci* **3**: 628-638

Florio SK, Prusti RK, Beavo JA (1996) Solubilization of membrane-bound rod phosphodiesterase by the rod phosphodiesterase recombinant delta subunit. *J Biol Chem* **271**: 24036-24047

Fotiadis D, Liang Y, Filipek S, Saperstein DA, Engel A, Palczewski K (2004) The G protein-coupled receptor rhodopsin in the native membrane. *FEBS Lett* **564**: 281-288

Fukata Y, Itoh TJ, Kimura T, Ménager C, Nishimura T, Shiromizu T, Watanabe H, Inagaki N, Iwamatsu A, Hotani H, Kaibuchi K (2002) CRMP-2 binds to tubulin heterodimers to promote microtubule assembly. *Nat Cell Biol* **4**: 583-591

Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, *et al* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141-147

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, *et al* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**: 631-636

Gharahdaghi F, Weinberg CR, Meagher DA, Imai BS, Mische SM (1999) Mass spectrometric identification of proteins from silver-stained polyacrylamide gel: a method for the removal of silver ions to enhance sensitivity. *Electrophoresis* **20**: 601-605

Gloeckner CJ, Boldt K, Ueffing M (2009) Strep/FLAG tandem affinity purification (SF-TAP) to study protein interactions. *Curr Protoc Protein Sci* **19**: Unit19

Gray SM, Kelly S, Robles LJ (2008) Rho signaling mediates cytoskeletal re-arrangements in octopus photoreceptors. *Am Malacol Bull* **26**: 19-26

Hall C, Brown M, Jacobs T, Ferrari G, Cann N, Teo M, Monfries C, Lim L (2001) Collapsin response mediator protein switches RhoA and Rac1 morphology in N1E-115 neuroblastoma cells and is regulated by Rho kinase. *J Biol Chem* **276**: 43482-43486

Hamer RD, Nicholas SC, Tranchina D, Lamb TD, Jarvinen JL (2005) Toward a unified model of vertebrate rod phototransduction. *Vis Neurosci* **22**: 417-436

Hanzal-Bayer M, Renault L, Roversi P, Wittinghofer A, Hillig RC (2002) The complex of Arl2-GTP and PDE delta: from structure to function. *EMBO J* **21**: 2095-2106

Henrick K, Thornton JM (1998) PQS: a protein quaternary structure file server. *Trends Biochem Sci* **9**: 358-361

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, *et al* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**: 180-183

Hofmann KP, Spahn CM, Heinrich R, Heinemann U (2006) Building functional modules from molecular interactions. *Trends Biochem Sci* **31**: 497-508

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**: 4569-4574

Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stümpflen V, Salwinski L, Nerothin J, Cerami E, *et al* (2007) Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol* **5**: 44

Kiel C, Serrano L (2006) The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes. *J Mol Biol* **355**: 821-844

Kiel C, Beltrao P, Serrano L (2008) Analyzing protein interaction networks using structural information. *Annu Rev Biochem* **77**: 415-441

Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**: 1938-1941

Krishnan A, Duda T, Pertzev A, Kobayashi M, Takamatsu K, Sharma RK (2009) Hippocalcin, new Ca(2+) sensor of a ROS-GC subfamily member, ONE-GC, membrane guanylate cyclase transduction system. *Mol Cell Biochem* **325**: 1-14

Kühn H (1978) Light-regulated binding of rhodopsin kinase and other proteins to cattle photoreceptor membranes. *Biochemistry* **17**: 4389-4395

Kühn H, Hall SW, Wilden U (1984). Light-induced binding of 48-kDa protein to photoreceptor membranes is highly enhanced by phosphorylation of rhodopsin. *FEBS Lett* **176**: 473-478

Kwok MC, Holopainen JM, Molday LL, Foster LJ, Molday RS (2008) Proteomics of photoreceptor outer segments identifies a subset of SNARE and Rab proteins

implicated in membrane vesicle trafficking and fusion. *Mol Cell Proteomics* **7.6**: 1053-1066

Lamb TD, Pugh EN Jr (2004) Dark adaptation and the retinoid cycle of vision. *Prog Retin Eye Res* **23**: 307-830

Liang Y, Fotiadis D, Filipek S, Saperstein DA, Palczewski K, Engel A (2003) Organization of the G protein-coupled receptors rhodopsin and opsin in native membranes. *J Biol Chem* **278**: 21655-21662

Liu BP, Strittmatter SM (2001) Semaphorin-mediated axonal guidance via Rho-related G proteins. *Curr Opin Cell Biol* **13**: 619-626

Liu Q, Tan G, Levenkova N, Li T, Pugh EN Jr, Rux JJ, Speicher DW, Pierce EA (2007) The proteome of the mouse photoreceptor sensory cilium complex. *Mol Cell Proteomics* **6.8**: 1299-1317

Marzesco AM, Galli T, Louvard D, Zahraoui A (1998) The rod cGMP phosphodiesterase delta subunit dissociates the small GTPase Rab13 from membranes. *J Biol Chem* **273**: 22340-22345

Mazelova J, Astuto-Gribble L, Inoue H, Tam BM, Schonteich E, Prekeris R, Moritz OL, Randazzo PA, Deretic D (2009) Ciliary targeting motif VxPx directs assembly of a trafficking module through Arf4. *EMBO J* **28**: 183-912

Mitchell R, McCulloch D, Lutz E, Johnson M, MacKenzie C, Fennell M, Fink G, Zhou W, Sealfon SC (1998) Rhodopsin-family receptors associate with small G proteins to activate phospholipase D. *Nature* **392**: 411-414

Molday RS, Hicks D, Molday L (1987) Peripherin. A rim-specific membrane protein of rod outer segment discs. *Invest Ophthalmol Vis Sci* **28**: 50-61

Mueller B, Eichacker L (1999) Assembly of the D1 Precursor in Monomeric Photosystem II Reaction Center Precomplexes Precedes Chlorophyll a–Triggered Accumulation of Reaction Center II in Barley Etioplasts. *The Plant Cell* **11**: 2365-2377

Nair KS, Hanson SM, Kennedy MJ, Hurley JB, Gurevich VV, Slepak VZ (2004) Direct binding of visual arrestin to microtubules determines the differential subcellular localization of its splice variants in rod photoreceptors. *J Biol Chem* **279**: 41240-41248

Nickell S, Park PS, Baumeister W, Palczewski K (2007) Three-dimensional architecture of murine rod outer segments determined by cryoelectron tomography. *J Cell Biol* **177**: 917-925

Nijtmans LG, Henderson NS, Holt IJ (2002) Blue Native electrophoresis to study mitochondrial and other protein complexes. *Methods* **26**: 327-334

Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**: 635-648

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289**: 739-745

Palczewski K (2006) G protein-coupled receptor rhodopsin. *Annu Rev Biochem* **75**: 743-767

Panfoli I, Calzia D, Bianchini P, Ravera S, Diaspro A, Candiano G, Bachi A, Monticone M, Aluigi MG, Barabino S, Calabria G, Rolando M, Tacchetti C, Morelli A, Pepe IM. (2009) Evidence for aerobic metabolism in retinal rod outer segment disks. *Int J Biochem Cell Biol* **41**: 2555-2565

Papermaster DS, Dreyer WJ (1974) Rhodopsin content in the outer segment membranes of bovine and frog retinal rods. *Biochemistry* **13**: 2438-2444

Paquet-Durand F, Beck S, Michalakis S, Goldmann T, Huber G, Muhlfriedel R, Trifunovic D, Fischer M D, Fahl E., Duetsch G, Becirovic E, Wolfrum U, van Veen T, Biel M, Tanimoto N, and Seeliger MW (2010). A key role for cyclic-nucleotide gated (CNG) channels in cGMP-related retinitis pigmentosa**.** *Hum Mol Genet* (Epub ahead of print)

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**: 3551-3567

Plantner JJ, Kean EL (1976) Carbohydrate composition of bovine rhodopsin. *J Biol Chem* **251**: 1548-1552

Poetsch A, Molday LL, Molday RS (2001) The cGMP-gated channel and related glutamic acid-rich proteins interact with peripherin-2 at the rim region of rod photoreceptor disc membranes. *J Biol Chem* **276**: 48009-48016

Reidel B, Goldmann T, Giessl A, Wolfrum U (2008) The translocation of signaling molecules in dark adapting mammalian rod photoreceptor cells is dependent on the cytoskeleton. *Cell Motil Cytoskeleton* **65**: 785-800

Ridge KD, Abdulaev NG, Sousa M, Palczewski K (2003) Phototransduction: crystal clear. *Trends Biochem Sci* **28**: 479-487

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, *et al* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**: 1173-1178

Santonico E, Castagnoli L, Cesareni G (2005) Methods to reveal domain networks. *Drug Discov Today* **10**: 1111-1117

Schägger H, von Jagow G (1991) Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Anal Biochem* **199**: 223-231

Scheffzek K, Stephan I, Jensen ON, Illenberger D, Gierschik P (2000) The Rac-RhoGDI complex and the structural basis for the regulation of Rho proteins by RhoGDI. *Nat Struct Biol* **7**: 122-126

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498-2504

Smith HG Jr, Stubbs GW, Litman BJ (1975) The isolation and purification of osmotically intact discs from retinal rod outer segments. *Exp Eye Res* **20**: 211-217

Stein A, Russell RB, Aloy P (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res* **33**: D413-D417

Stein A, Panjkovich A, Aloy P (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res* **37**: D300-D304

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N,

Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, *et al* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**: 957-968

Swiatek-de Lange M, Müller B, Ueffing M (2008) Native Fractionation: Isolation of Native Membrane-Bound Protein Complexes from Porcine Rod Outer Segments Using Isopycnic Density Gradient Centrifugation. *Methods in Molecular Biology* **484**: 161-175

Veltel S, Wittinghofer A (2009) RPGR and RP2: targets for the treatment of X-linked retinitis pigmentosa? *Expert Opin Ther Targets***13**: 1239-1251

Walikonis RS, Oguni A, Khorosheva EM, Jeng CJ, Asuncion FJ, Kennedy MB (2001). Densin-180 forms a ternary complex with the (alpha)-subunit of Ca2+/calmodulin-dependent protein kinase II and (alpha)-actinin. *J Neurosci* **21**: 423-433

Wensel TG (2008) Signal transducing membrane complexes of photoreceptor outer segments. *Vision Res* **48**: 2052-2061

Wenzel A, Grimm C, Samardzija M, Remé CE (2005) Molecular mechanisms of light-induced photoreceptor apoptosis and neuroprotection for retinal degeneration. *Prog Retin Eye Res* **24**: 275-306

Wessel D, Flügge UI (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **138**: 141-143

Wieland T, Ulibarri I, Aktories K, Gierschik P, Jakobs KH (1990a) Interaction of small G proteins with photoexcited rhodopsin. *FEBS Lett* **263**: 195-198

Wieland T, Ulibarri I, Gierschik P, Hall A, Aktories K, Jakobs KH (1990b) Interaction of recombinant rho A GTP-binding proteins with photoexcited rhodopsin. *FEBS Lett* **274**: 111-114

Wilden U, Wust E, Weyand I, Kuhn H (1986) Rapid affinity purification of retinal arrestin (48 kDa protein) via its light-dependent binding to phosphorylated rhodopsin. *FEBS Lett* **207**: 292-295

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) MINT: a Molecular INTeraction database. *FEBS Lett* **513**: 135-140

Zhang H, Liu XH, Zhang K, Chen CK, Frederick JM, Prestwich GD, Baehr W (2004) Photoreceptor cGMP phosphodiesterase delta subunit (PDEdelta) functions as a prenyl-binding protein. *J Biol Chem* **279**: 407-413

Zischka H, Gloeckner CJ, Klein C, Willmann S, Swiatek-de Lange M, Ueffing M (2004) Improved mass spectrometric identification of gel-separated hydrophobic membrane proteins after sodium dodecyl sulfate removal by ion-pair extraction. *Proteomics* **4**: 3776-3782

Figure legends

Figure 1 Proteomic description of the retina ROS inventory and GO analysis. (**A**) Schematic model of a rod photoreceptor cell (left) and its corresponding location within the retina (depicted in the micrograph to the right). Segments labeled in the model are: rod outer segment (ROS) with enclosed stacks of discs membranes containing the visual pigment molecules rhodopsin; connecting cilium (CC); rod inner segment (RIS) containing mitochondria, Golgi, and ER membranes, and vesicles in which opsin molecules are assembled before transported to the outer segment; and the cell body containing the nucleus and a synaptic termini, where neurotransmission to second-order neurons occurs. The micrograph depicts the vertical porcine retina with its cytoarchitectural organization labeled as: photoreceptor outer segments (OS); the outer nuclear layer (ONL) containing cell bodies of rods and cones; the outer plexiform layer (OPL); the inner nuclear layer (INL); the inner plexiform layer (IPL), and the ganglion cell layer (GCL). The retinal pigment epithelium (RPE) is localized above the photoreceptor cell layer (for details, see http://webvision.med.utah.edu). Retinal cells nuclei were stained with DAPI (magnification 40×). Insets show micrographs of the OS immunolabeled with anti-rhodopsin with an FITC-conjugated secondary antibody (magnification 40×; top inset), and of the OS preparation (magnification 40×; bottom inset). (**B**) Comparison of different proteomic data sets determined in ROS, based on proteins and the protein overlap identified in the proteomic analysis from this work and that of Kwok *et al* (2008). The union of the two datasets was defined as the *initial experimental ROS proteome*. (**C**) Functional modules and GO analyses of the filtered *core ROS proteome*. By performing an automatic and a manual gene ontology (GO) search (based on the UniProt and KEGG databases), we characterized the 355 proteins (see Supplementary Table S2) to be involved in: vision, signaling, transport, and channels (56), disc structure and morphology (7), housekeeping functions (73), cytoskeleton and polarity (67), vesicle, structure, and trafficking (60), and metabolism (92). Sub-modules/sub-functions of the GO terms are indicated as described in Supplementary Table S2 (1A, phototransduction/ channels (33); 1B, retinol recycling (5); 1C, calcium signaling (18); 2A, disk morphology (2); 2B, link to ECM (5); 3A, protein folding (8); 3B, chaperones/ heat shock (25); 3C, ubiquitination/degradation/proteasome (10); 3D,

scaffolds/adaptor proteins (7); 3E, oxidative stress/cell redox homoestasis (9); 3F, apoptosis (2); 3G, others (2); 3H, signaling (10); 4A, regulation of cytoskeleton (34); 4B, cytoskeleton proteins (21); 4C, motor proteins (7); 4D, protein transport (1); 4E, axon guidance (4); 5A, endocytosis (10); 5B, exocytosis (8); 5C, Golgi endosome (11); 5D, vesicle transport/fusion (12); 5E, Golgi/ER/trafficking (19); 6A, glycolysis (20); 6B, tricarboxylic acid (5); 6C, ATP synthesis (25); 6D, lipid/fatty acids metabolism (9); 6E, amino acid metabolism (9); 6F, one-carbon metabolism (4); 6G, nucleotide metabolism (6); 6H, glucose/lipid/phosphate/amino acid/ion transport (8); 6I, pentose phosphate shunt (1); 6J, mevalonate (1); and 6K, others (4).

Figure 2 Experimental and computational workflow. The flow charts of experimental (yellow boxes) and bioinformatic (green boxes) methods used in this work are shown. The initial ROS proteome was generated based on the union of proteins identified in bovine ROS in this work and those from a proteomic analysis of porcine ROS (Kwok *et al*, 2008). After filtering, a high-confidence ROS proteome was defined. A static ROS interactome was compiled by literature mining. In addition, new experiments were performed in ROS in this work (co-sedimentation and co-immunoprecipitation). Further, we performed structural analyses and homology modeling, to distinguish between compatible and mutually exclusive interactions. This enabled us to break the network of nodes and edges into functional machines or sub-networks and modules. The comprehensive multiscale network highlights new predicted links and functions. Lastly, disease-associated genes were identified and modeled into available structures.

Figure 3 The high-confidence ROS interactome and the high-confidence binary ROS interactome. (**A**) The high-confidence ROS interactome. The 660 higher confidence interactions of the ROS interactome are listed (Supplementary Table S6). The size of the nodes indicates the number of interaction partners for a given protein (of >10 or >20). Edges with binary evidence are indicated with blue, while edges supported by more than one piece of evidence are indicated in grey. Proteins are colored according to their function. (**B**) The high-confidence binary ROS interactome. Modules and sub-modules are shown, and only the interactions of proteins from two different modules are indicated

(see Supplementary Material 2). The number of proteins implicated in diseases in each category is indicated.

Figure 4 Structural coverage of the core vision pathway and its links to other functional modules. The published core pathway (Dell'Orco *et al*, 2009) was extended using evidence from our high-confidence network. Outputs to different functional cellular processes emanating from the proteins in the pathway are indicated, and the available structures are displayed by ribbon representation (see the main text, and Supplementary Material 2). Proteins are colored according to their function.

Figure 5 Graphical representation of experiments performed in this work and its comparison with interactions described in the literature. Protein complexes that were obtained using Rac1, RhoA, or Rac1 as the bait protein are displayed within orange, blue, and yellow circles, respectively (see legend). The Rac1 and RhoA complexes were identified by Western blot, and the Rac1 complex, by Orbitrap. The overlap of the three circles indicates the proteins that were identified in the same complex in one of the three experiments. Connecting lines between proteins indicate either binary or co-immunoprecipitation interactions from the literature, or from BN-PAGE or co-sedimentation interactions as determined in this work. Proteins are colored according to their function.

Figure 6 Immunohistochemical analyses of porcine retina. Cryostat sections of the retina were stained with primary antibodies (red) against indicated proteins, and nuclei were counterstained (blue). The images on the left were taken from the outer retina [outer segments (OS), inner segments (IS), outer nuclear layer (ONL) and outer plexiform layer (OPL)]. Images in the middle are an overlay of antibody staining, nuclei staining, and DIC optics (Nomarski). Images on the right were taken with higher magnification, to focus on the OS and IS. All indicated proteins were unambigiously identified as constituents of ROS. Control sections without primary antibodies showed no staining (Supplementary Figure S5).

Figure 7 Experimental evidence that PDEδ acts as a GDI for Rac1 in ROS. (**A**) PDEδ and Rac1 colocalize in ROS in native protein complexes. After solubilization with β-dodecylmaltoside, native ROS protein complexes from soluble and membranous fractions of light- and dark-adapted ROS were separated by BN-PAGE. Components of the native protein complexes were separated by SDS-PAGE for second-dimension electrophoresis. Western blots with anti-Rac1 and anti-PDEδ antibodies showed that PDEδ and Rac1 colocalized but were in different complexes in ROS depending on the dark-adapted state of the retina. Colocalization of PDEδ and Rac1 seemed to be stronger in the dark-adapted state, where both proteins colocalized to the soluble and membranous fractions. In light-adapted ROS, colocalization of PDEδ and Rac1 was detected only in the membranous fraction but not in the soluble fraction. (**B**) In vitro solubilization of Rac1 GTPase from light- and dark-adapted ROS membranes. Membranes isolated from light- or dark-adapted ROS were incubated for 1 h at 37°C with different amounts of recombinant human PDEδ (rhPDEδ) or buffer alone, and the unsolubilized material was recovered by ultracentrifugation. Immunoblots with anti-Rac1 or anti-PDEδ antibodies showed that PDEδ solubilizes Rac1 from ROS membranes in a dose-dependent manner. Since it has been previously determined that PDEδ solubilizes PDEδ from ROS membranes in a dose-dependent manner (Florio *et al*, 1996), this was used here to demonstrate the functional activity of the rhPDEδ protein.

Figure 8 Network representations distinguishing mutually exclusive from compatible interactions, based on structural information. All protein-protein interactions for which structural information was available (Supplementary Table S4), and for which structural superimpositions were performed (Supplementary Figure S6), are represented here. Mutually exclusive complexes are indicated with "XOR", and compatible interactions are indicated with "AND". Proteins are colored according to their function (see Figure 3B).

Figure 1

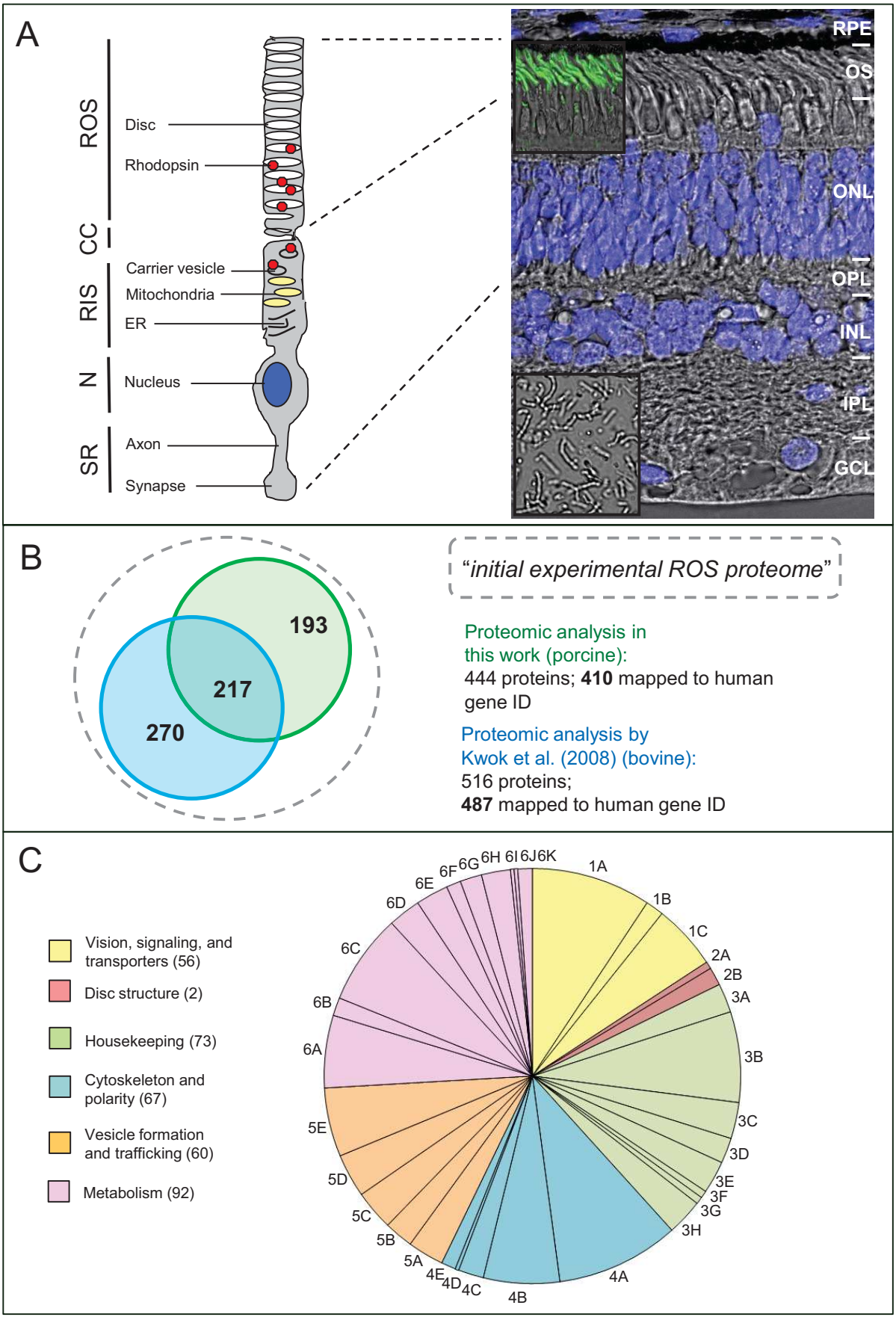

A
ROS
CC
RIS
N
SR
Disc
Rhodopsin
Carrier vesicle
Mitochondria
ER
Nucleus
Axon
Synapse
RPE
OS
ONL
OPL
INL
IPL
GCL
B
193
217
270
"initial experimental ROS proteome"
Proteomic analysis in this work (porcine):
444 proteins; 410 mapped to human gene ID
Proteomic analysis by Kwok et al. (2008) (bovine):
516 proteins;
487 mapped to human gene ID
C
Vision, signaling, and transporters (56)
Disc structure (2)
Housekeeping (73)
Cytoskeleton and polarity (67)
Vesicle formation and trafficking (60)
Metabolism (92)
1A
1B
1C
2A
2B
3A
3B
3C
3D
3E
3F
3G
3H
4A
4B
4C
4D
4E
5A
5B
5C
5D
5E
6A
6B
6C
6D
6E
6F
6G
6H
6I
6J
6K

Figure 2

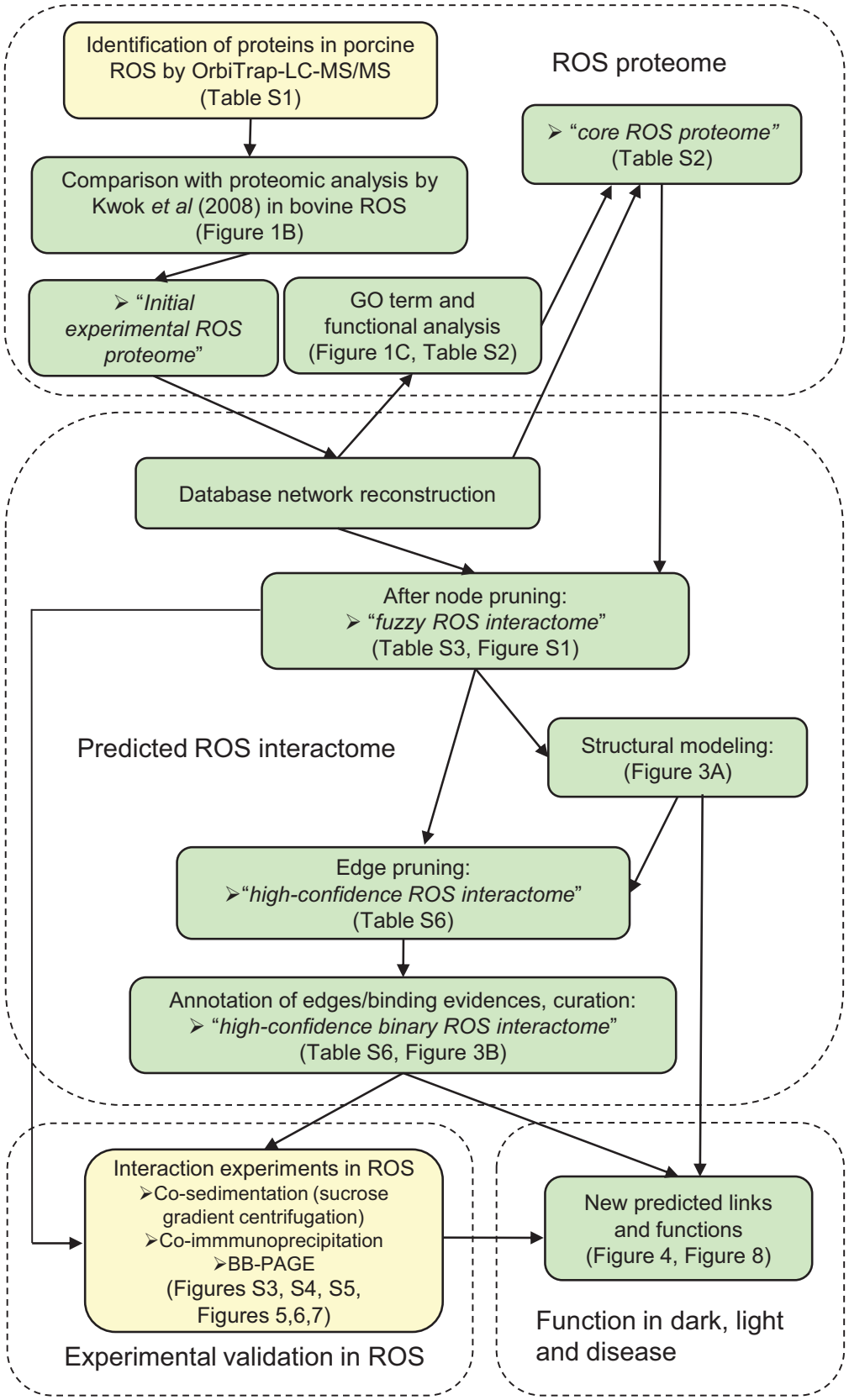

Identification of proteins in porcine ROS by OrbiTrap-LC-MS/MS (Table S1)
ROS proteome
➢ "core ROS proteome" (Table S2)
Comparison with proteomic analysis by Kwok et al (2008) in bovine ROS (Figure 1B)
➢ "Initial experimental ROS proteome"
GO term and functional analysis (Figure 1C, Table S2)
Database network reconstruction
After node pruning:
➢ "fuzzy ROS interactome"
(Table S3, Figure S1)
Predicted ROS interactome
Structural modeling:
(Figure 3A)
Edge pruning:
➢"high-confidence ROS interactome"
(Table S6)
Annotation of edges/binding evidences, curation:
➢ "high-confidence binary ROS interactome"
(Table S6, Figure 3B)
Interaction experiments in ROS
➢Co-sedimentation (sucrose gradient centrifugation)
➢Co-immmunoprecipitation
➢BB-PAGE
(Figures S3, S4, S5, Figures 5,6,7)
Experimental validation in ROS
New predicted links and functions
(Figure 4, Figure 8)
Function in dark, light and disease

Figure 3

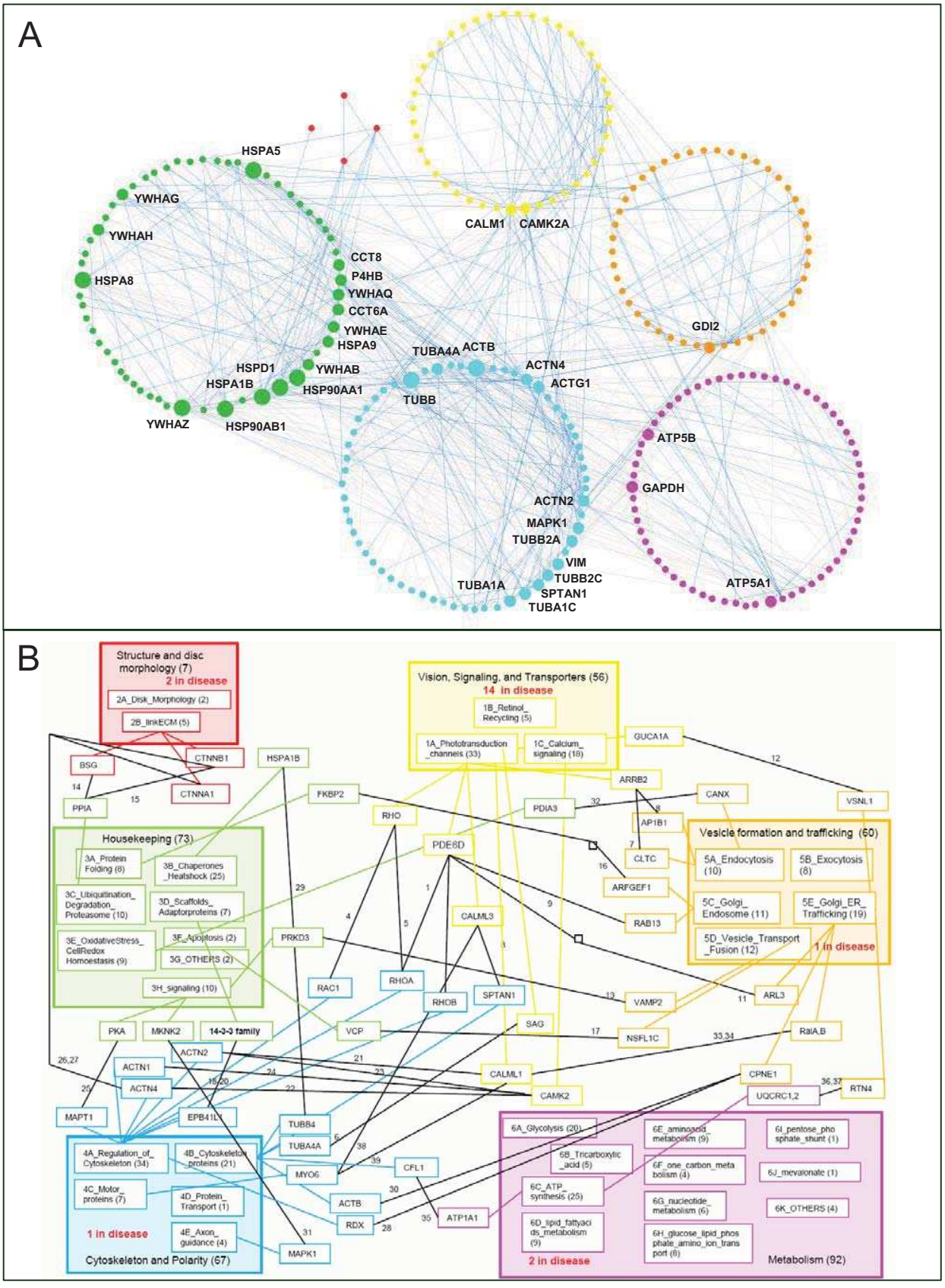

A
HSPA5
YWHAG
YWHAH
HSPA8
CCT8
P4HB
YWHAQ
CCT6A
YWHAE
HSPA9
HSPD1
YWHAB
HSPA1B
HSP90AA1
YWHAZ
HSP90AB1
CALM1
CAMK2A
GDI2
TUBA4A
ACTB
ACTN4
ACTG1
TUBB
ATP5B
GAPDH
ACTN2
MAPK1
TUBB2A
VIM
TUBB2C
TUBA1A
SPTAN1
TUBA1C
ATP5A1
B
Structure and disc morphology (7)
2 in disease
2A_Disk_Morphology (2)
2B_linkECM (5)
Vision, Signaling, and Transporters (56)
14 in disease
1B_Retinol_ Recycling (5)
1A_Phototransduction _channels (33)
1C_Calcium_ signaling (18)
Housekeeping (73)
3A_Protein Folding (8)
3B_Chaperones _Heatshock (25)
3C_Ubiquitination_ Degradation_ Proteasome (10)
3D_Scaffolds_ Adaptorproteins (7)
3E_OxidativeStress_ CellRedox Homoestasis (9)
3F_Apoptosis (2)
3G_OTHERS (2)
3H_signaling (10)
Vesicle formation and trafficking (60)
5A_Endocytosis (10)
5B_Exocytosis (8)
5C_Golgi_ Endosome (11)
5E_Golgi_ER_ Trafficking (19)
5D_Vesicle_Transport _Fusion (12)
1 in disease
Cytoskeleton and Polarity (67)
4A_Regulation_of_ Cytoskeleton (34)
4B_Cytoskeleton _proteins (21)
4C_Motor_ proteins (7)
4D_Protein_ Transport (1)
4E_Axon_ guidance (4)
1 in disease
Metabolism (92)
6A_Glycolysis (20)
6B_Tricarboxylic _acid (5)
6C_ATP_ synthesis (25)
6D_lipid_fattyaci ds_metabolism (9)
6E_aminoacid_ metabolism (9)
6F_one_carbon_meta bolism (4)
6G_nucleotide_ metabolism (6)
6H_glucose_lipid_phos phate_amino_ion_trans port (8)
6I_pentose_pho sphate_shunt (1)
6J_mevalonate (1)
6K_OTHERS (4)
2 in disease
BSG
CTNNB1
HSPA1B
PPIA
CTNNA1
FKBP2
RHO
PDIA3
ARRB2
CANX
VSNL1
AP1B1
PDE8D
CLTC
ARFGEF1
RAB13
CALML3
PRKD3
RAC1
RHOA
RHOB
SPTAN1
VAMP2
ARL3
PKA
MKNK2
14-3-3 family
VCP
SAG
NSFL1C
RalA,B
ACTN2
ACTN1
ACTN4
CALML1
CAMK2
CPNE1
UQCRC1,2
RTN4
MAPT1
EPB41L2
TUBB4
TUBA4A
MYO6
CFL1
ACTB
RDX
ATP1A1
MAPK1
GUCA1A
12
14
15
32
8
7
16
29
1
4
5
9
13
11
26,27
25
24
23
22
21
18,20
6
38
39
30
35
28
31
17
33,34
36,37

Figure 4

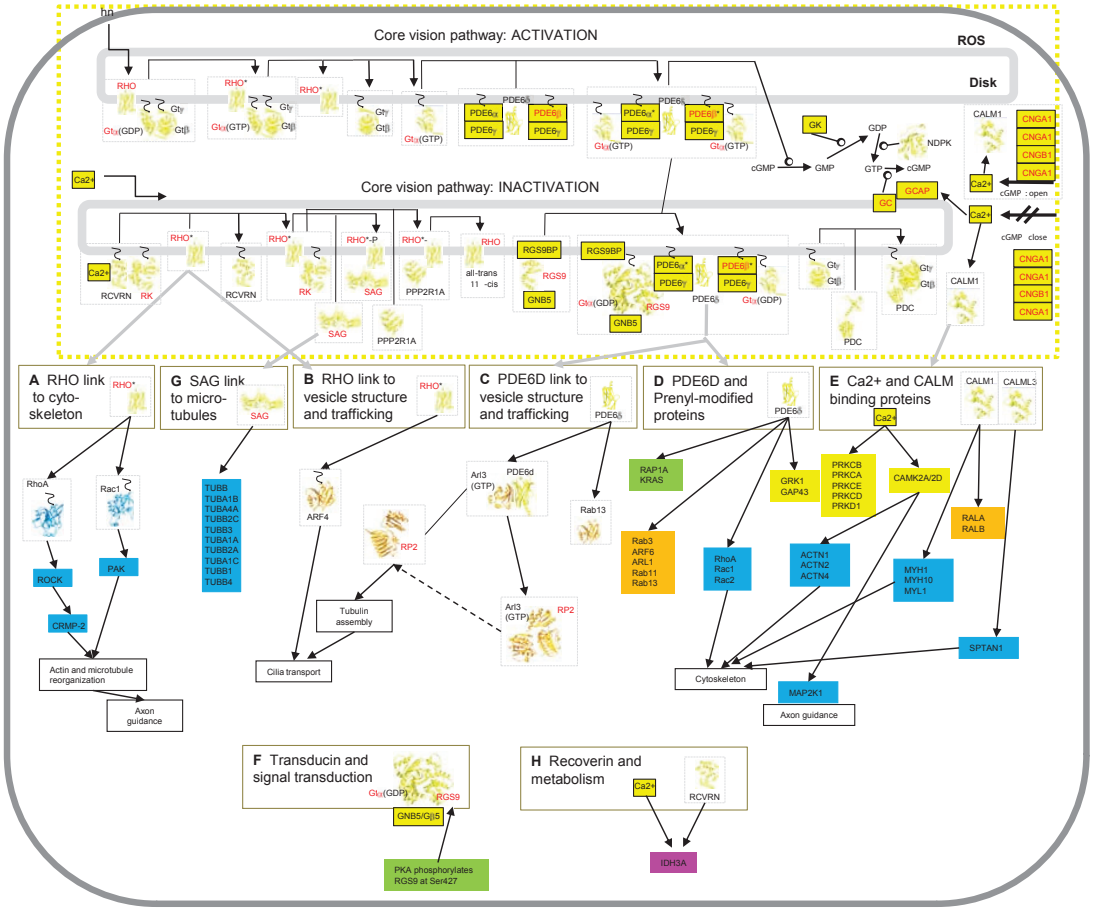

Core vision pathway: ACTIVATION
ROS
Disk
Core vision pathway: INACTIVATION
A RHO link to cyto-skeleton
G SAG link to micro-tubules
B RHO link to vesicle structure and trafficking
C PDE6D link to vesicle structure and trafficking
D PDE6D and Prenyl-modified proteins
E Ca2+ and CALM binding proteins
F Transducin and signal transduction
H Recoverin and metabolism
Actin and microtubule reorganization
Axon guidance
Cilia transport
Cytoskeleton
MAP2K1
PKA phosphorylates RGS9 at Ser427
IDH3A

Figure 5

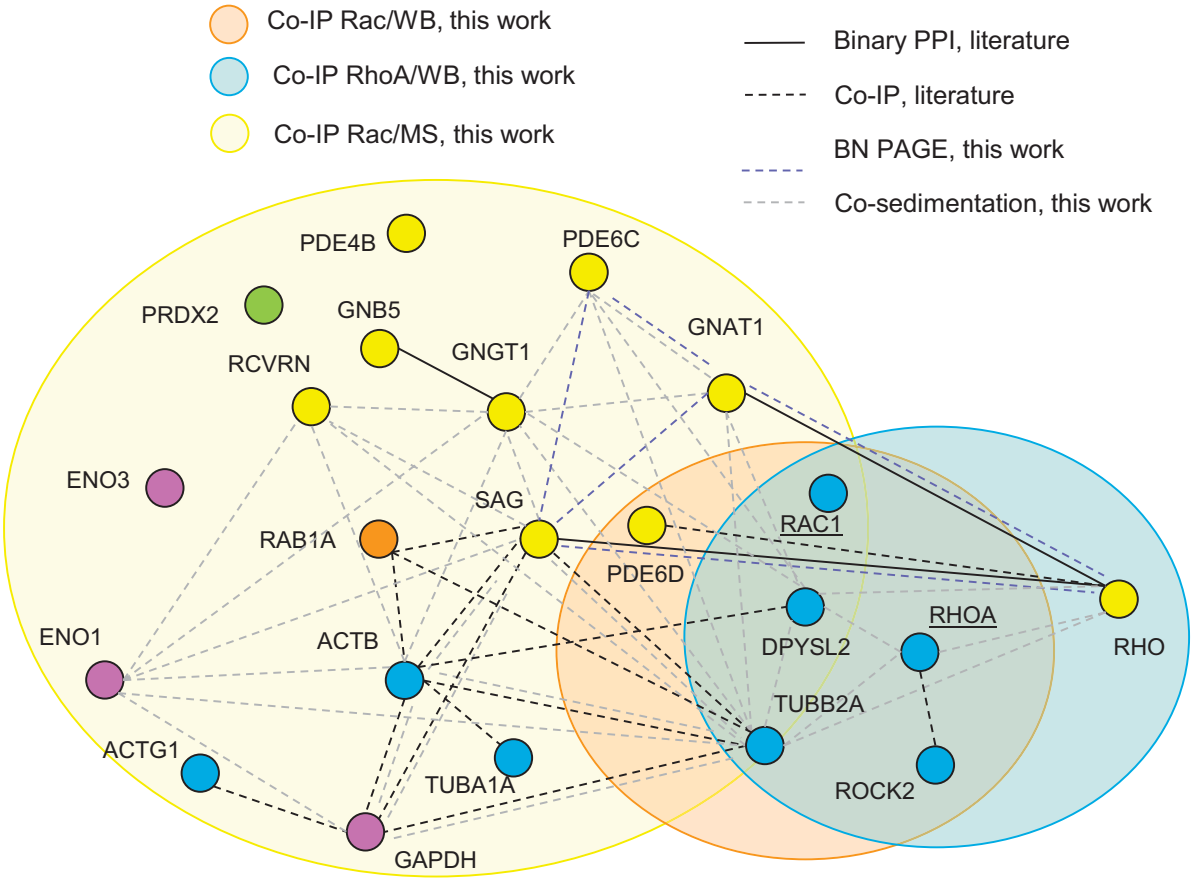

Co-IP Rac/WB, this work
Co-IP RhoA/WB, this work
Co-IP Rac/MS, this work
Binary PPI, literature
Co-IP, literature
BN PAGE, this work
Co-sedimentation, this work
PDE4B
PDE6C
PRDX2
GNB5
GNAT1
GNGT1
RCVRN
ENO3
SAG
RAC1
RAB1A
PDE6D
RHOA
DPYSL2
RHO
ENO1
ACTB
TUBB2A
ACTG1
TUBA1A
ROCK2
GAPDH

Figure 6

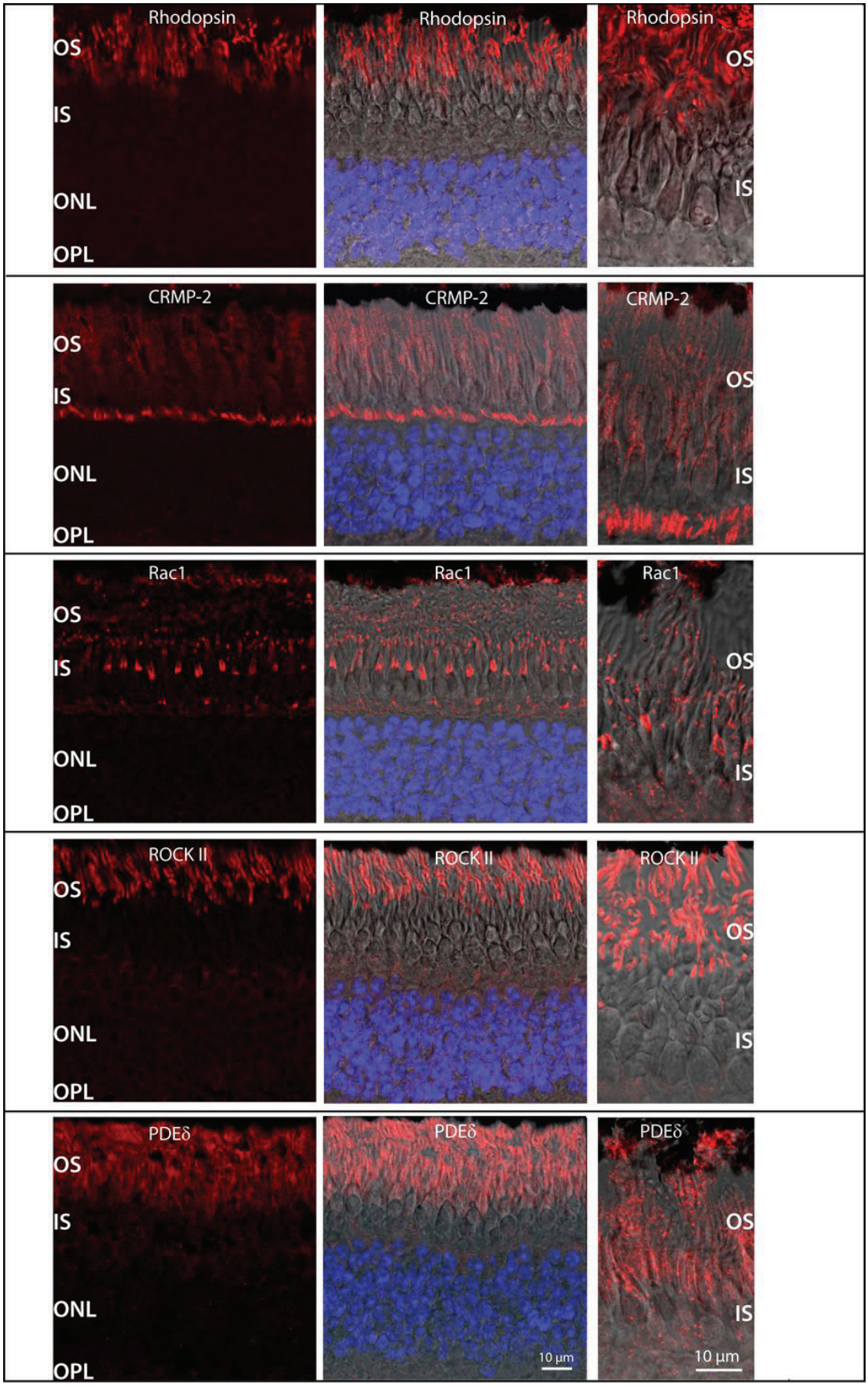

Rhodopsin
OS
IS
ONL
OPL
Rhodopsin
Rhodopsin
OS
IS
CRMP-2
OS
IS
ONL
OPL
CRMP-2
CRMP-2
OS
IS
Rac1
OS
IS
ONL
OPL
Rac1
Rac1
OS
IS
ROCK II
OS
IS
ONL
OPL
ROCK II
ROCK II
OS
IS
PDEδ
OS
IS
ONL
OPL
PDEδ
PDEδ
OS
IS
10 µm
10 µm

Figure 7

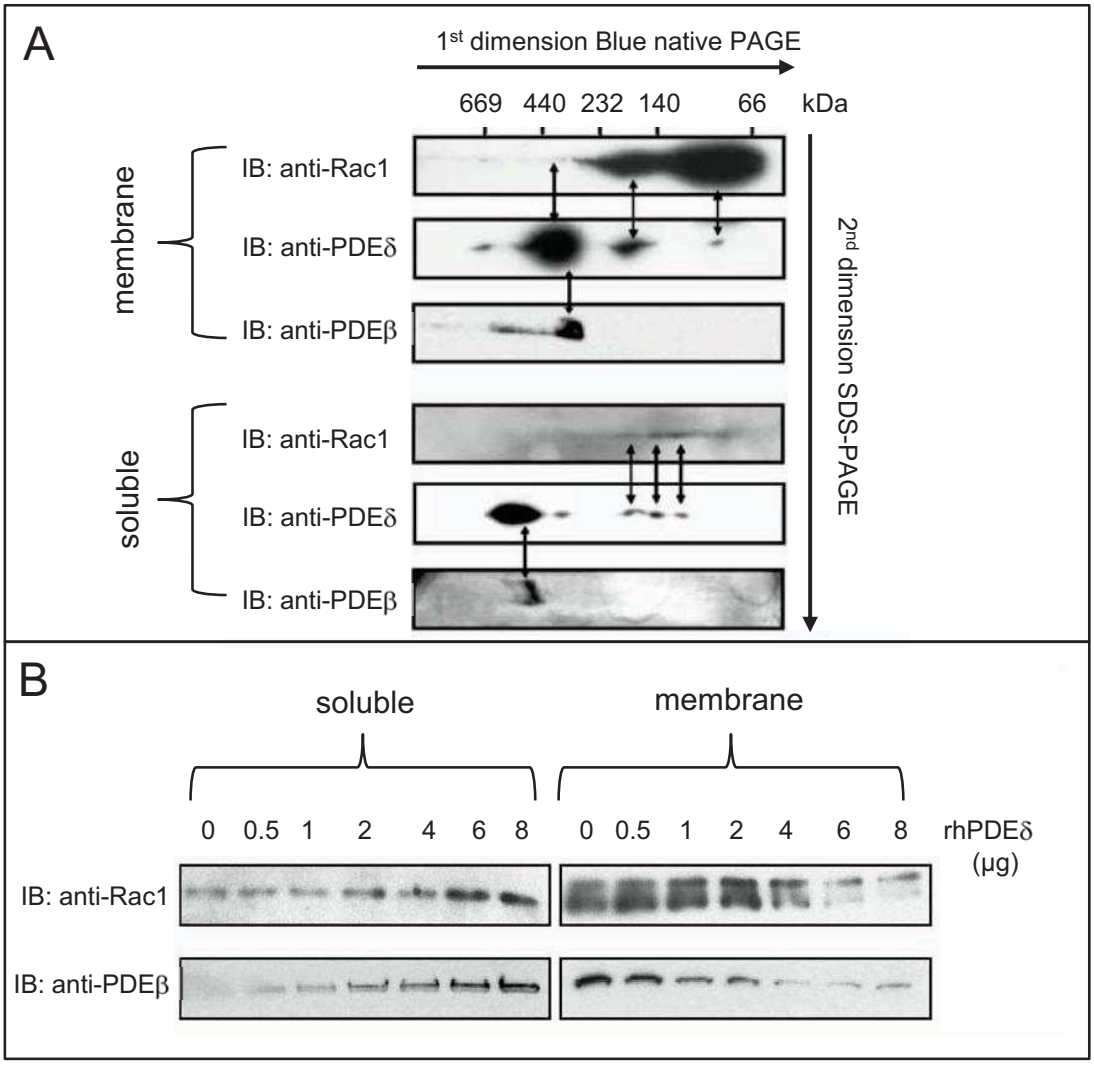
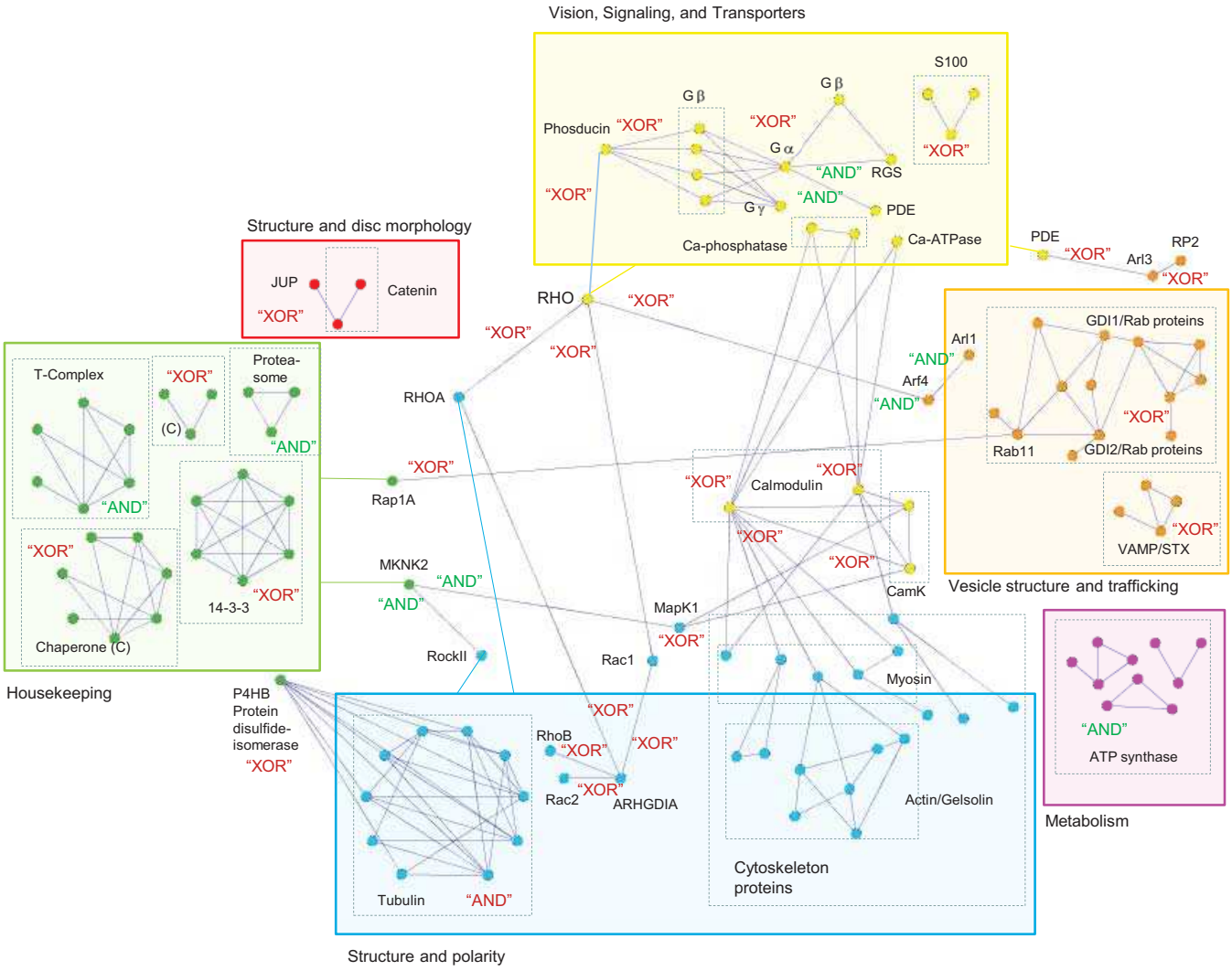
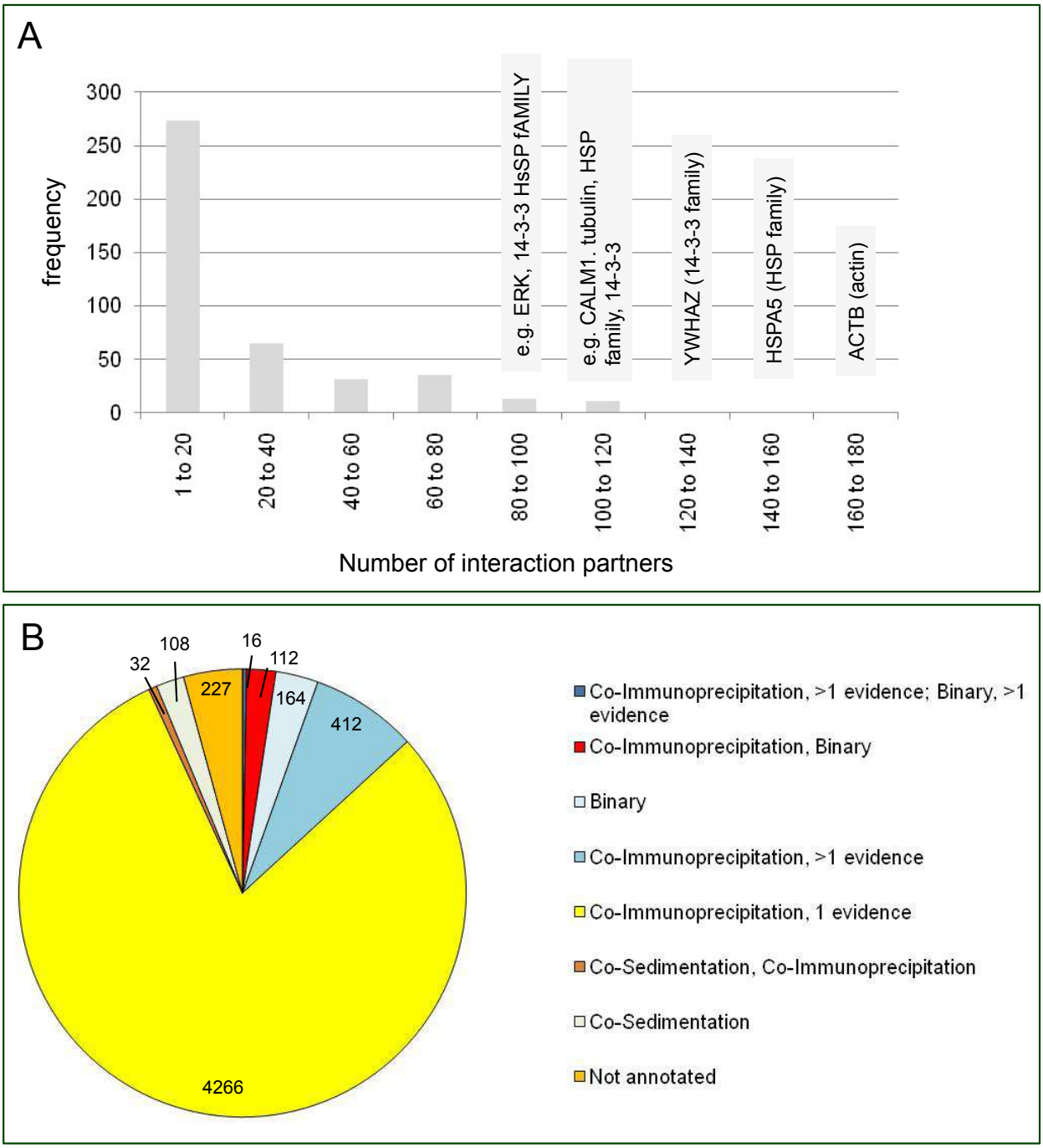

A
1st dimension Blue native PAGE
669 440 232 140 66 kDa
membrane
IB: anti-Rac1
IB: anti-PDEδ
IB: anti-PDEβ
soluble
IB: anti-Rac1
IB: anti-PDEδ
IB: anti-PDEβ
2nd dimension SDS-PAGE
B
soluble
membrane
0 0.5 1 2 4 6 8 0 0.5 1 2 4 6 8
rhPDEδ (μg)
IB: anti-Rac1
IB: anti-PDEβ

Figure 8

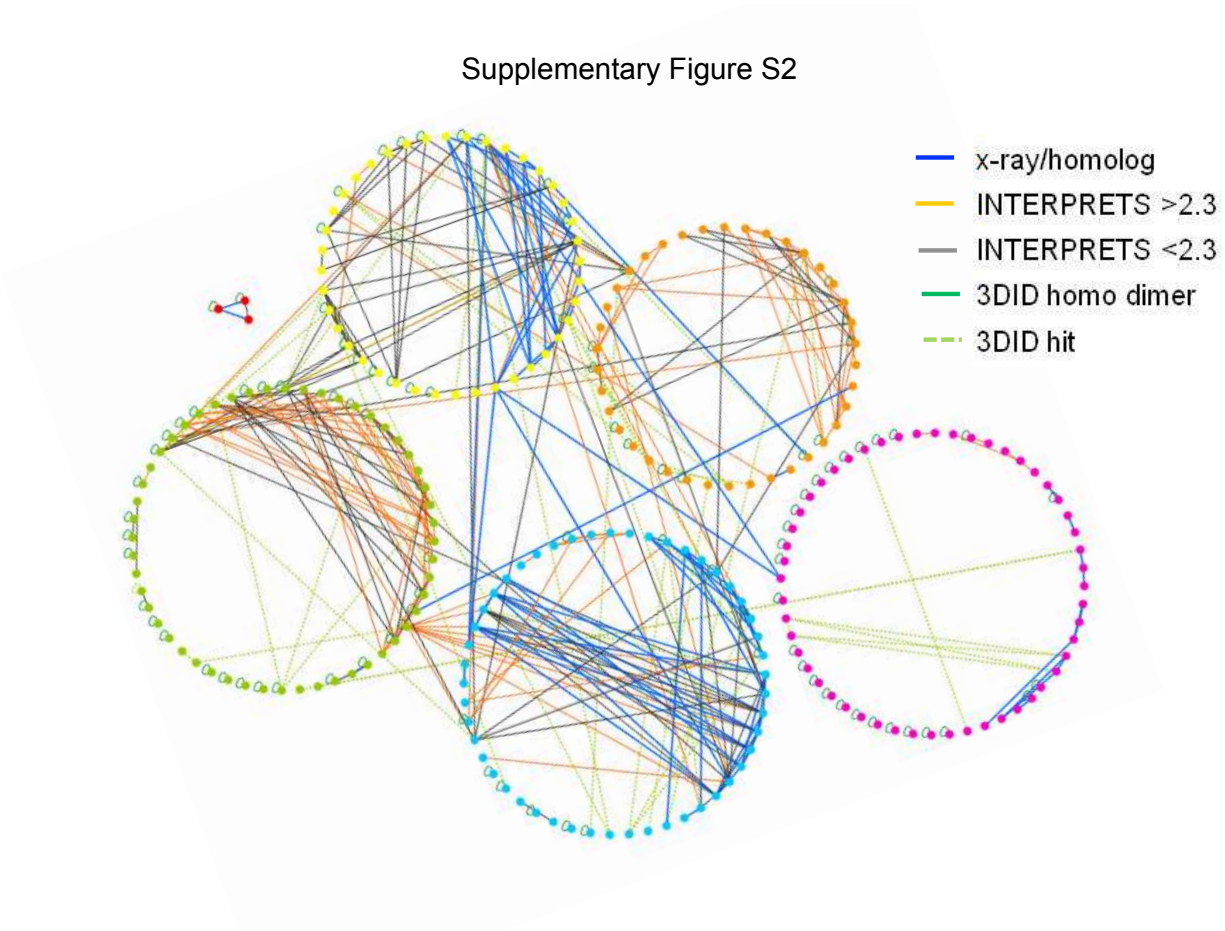

Vision, Signaling, and Transporters
S100
G β
G β
Phosducin
"XOR"
"XOR"
G α
"XOR"
"XOR"
"AND"
RGS
G γ
"AND"
PDE
Ca-phosphatase
Ca-ATPase
PDE
"XOR"
RP2
Arl3
"XOR"
Structure and disc morphology
JUP
Catenin
"XOR"
RHO
"XOR"
"XOR"
"XOR"
Arl1
"AND"
Arf4
"AND"
GDI1/Rab proteins
"XOR"
Rab11
GDI2/Rab proteins
VAMP/STX
"XOR"
Vesicle structure and trafficking
T-Complex
"XOR"
Protea-
some
(C)
"AND"
"AND"
RHOA
"XOR"
Rap1A
"XOR"
"XOR"
Calmodulin
"XOR"
"XOR"
"XOR"
CamK
"XOR"
14-3-3
"XOR"
Chaperone (C)
MKNK2
"AND"
"AND"
MapK1
"XOR"
RockII
Rac1
Myosin
Housekeeping
P4HB
Protein
disulfide-
isomerase
"XOR"
"XOR"
RhoB
"XOR"
"XOR"
"XOR"
Rac2
ARHGDIA
Actin/Gelsolin
"AND"
ATP synthase
Metabolism
Tubulin
"AND"
Cytoskeleton
proteins
Structure and polarity

Supplementary Figure S1



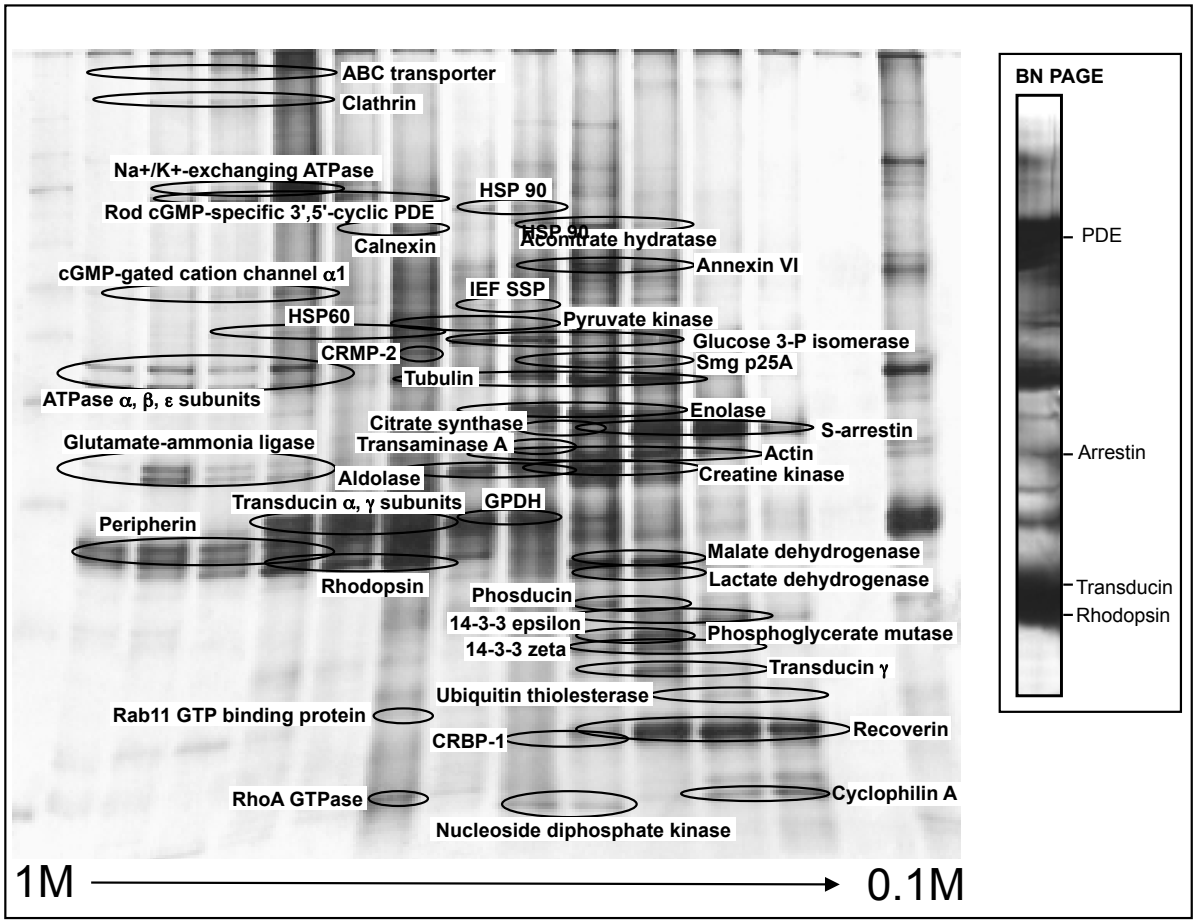
Supplementary Figure S1 Characteristics of the initial low confidence interactome ("fuzzy ROS interactome"). (**A**) Frequency of number of interaction partners for the 431 proteins in the interactome. All interactions and evidences are summarized in Supplementary Table S3. (**B**) Literature evidences for 5335 protein protein interactions among the 676 proteins of the initial ROS proteome. We divided the evidences for every PPI, as stored in the MINT database, into three categories, and their combinations: (i) very weak evidence for true binary interactions (Co-Sedimentation), (ii) weak evidence for binary interaction (Co-Immunoprecipitation), and (iii) true binary interactions (Binary). For details see Material and Methods.

Supplementary Figure S2



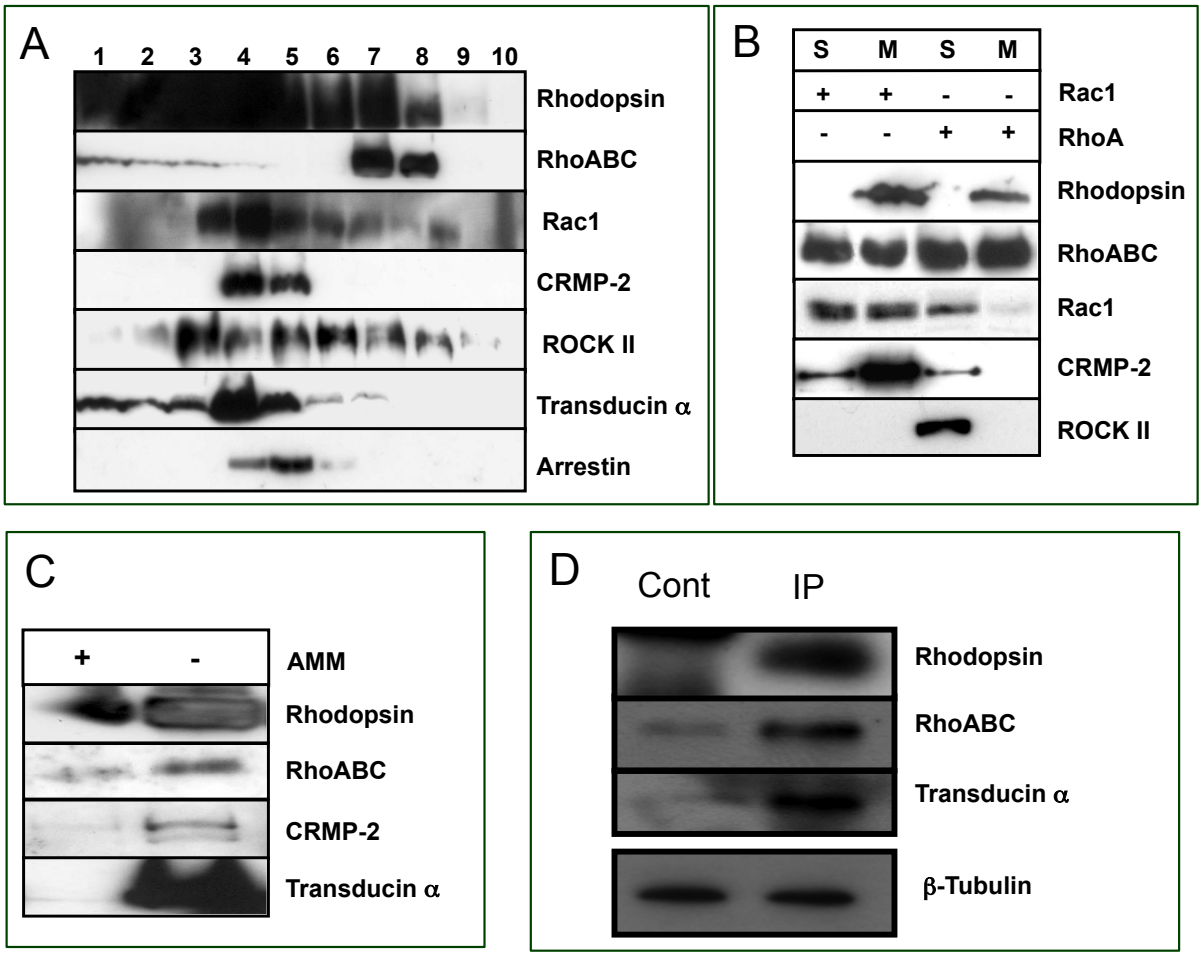
Supplementary Figure S2 Results from structural modeling. Summary of protein-protein interactions supported by structural modeling. In total 436 evidences were found, with 84 are from x-ray or close homolog structure, and 107 have a significant score when assessed using InterPreTS. A further 96 interactions have an insignificant score, 149 interactions could possibly be interact through the similar domain-domain based on the 3DID database, but they could not be scored using InterPreTS. (B) The high confidence ROS interactome. The 694 higher confidence interactions of the ROS interactome (Supplementary Table S6). Proteins are colored according to function.

Supplementary Figure S3



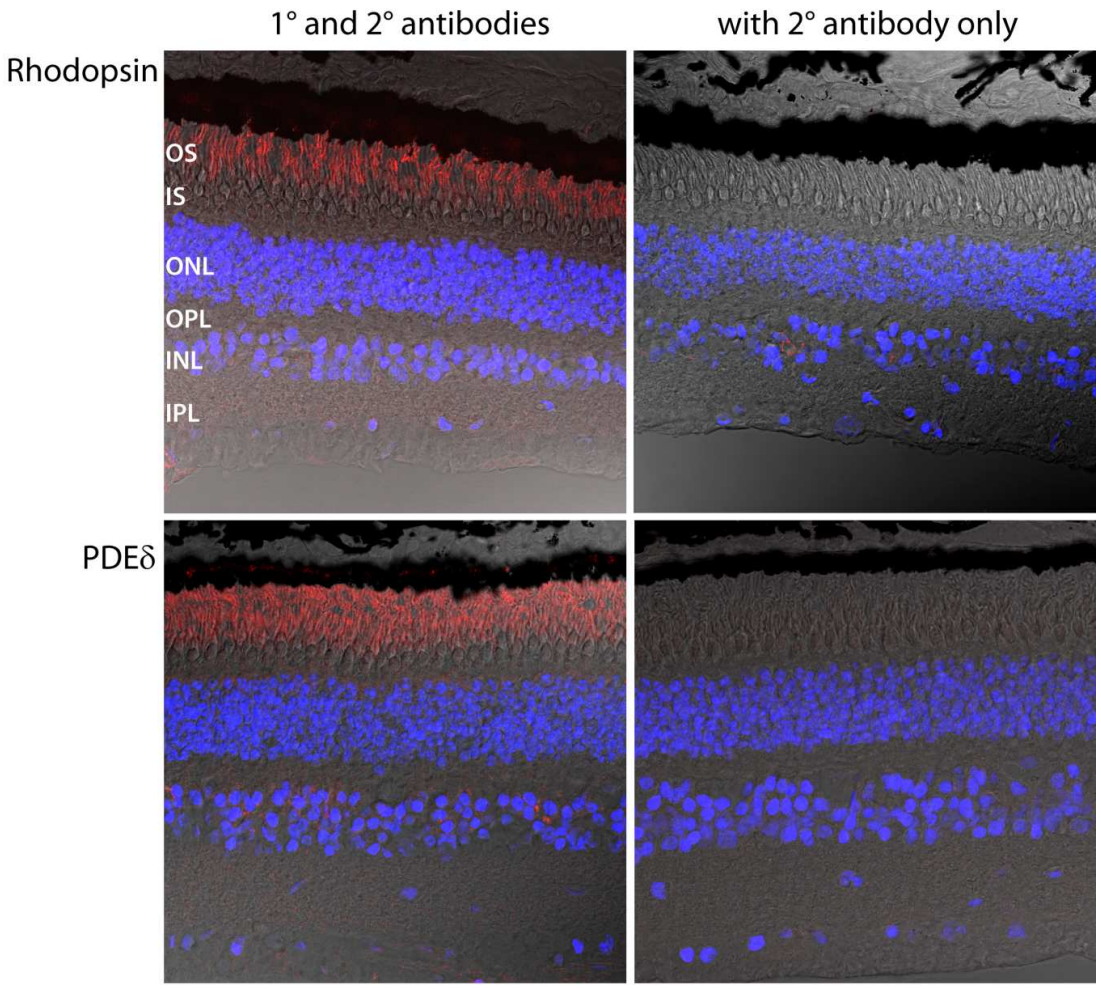
Supplementary Figure S3 Identification of native ROS membrane protein complexes by sucrose density gradient centrifugation and BN PAGE. Solubilized porcine ROS membrane proteins were subjected to sucrose density gradient centrifugation (0.1-1M). Fractions were collected from sucrose density gradient and analyzed by SDS–PAGE (9-15%). Inset: Section of BN PAGE (4-12%) of solubilized porcine ROS membrane proteins. Silver-stained SDS-PAGE bands were cut out and proteins were identified by mass spectrometry. Proteins co-migrating in the same fraction could belong to the same native protein complex. For a comprehensive list see Supplementary Table S6.

Supplementary Figure S4



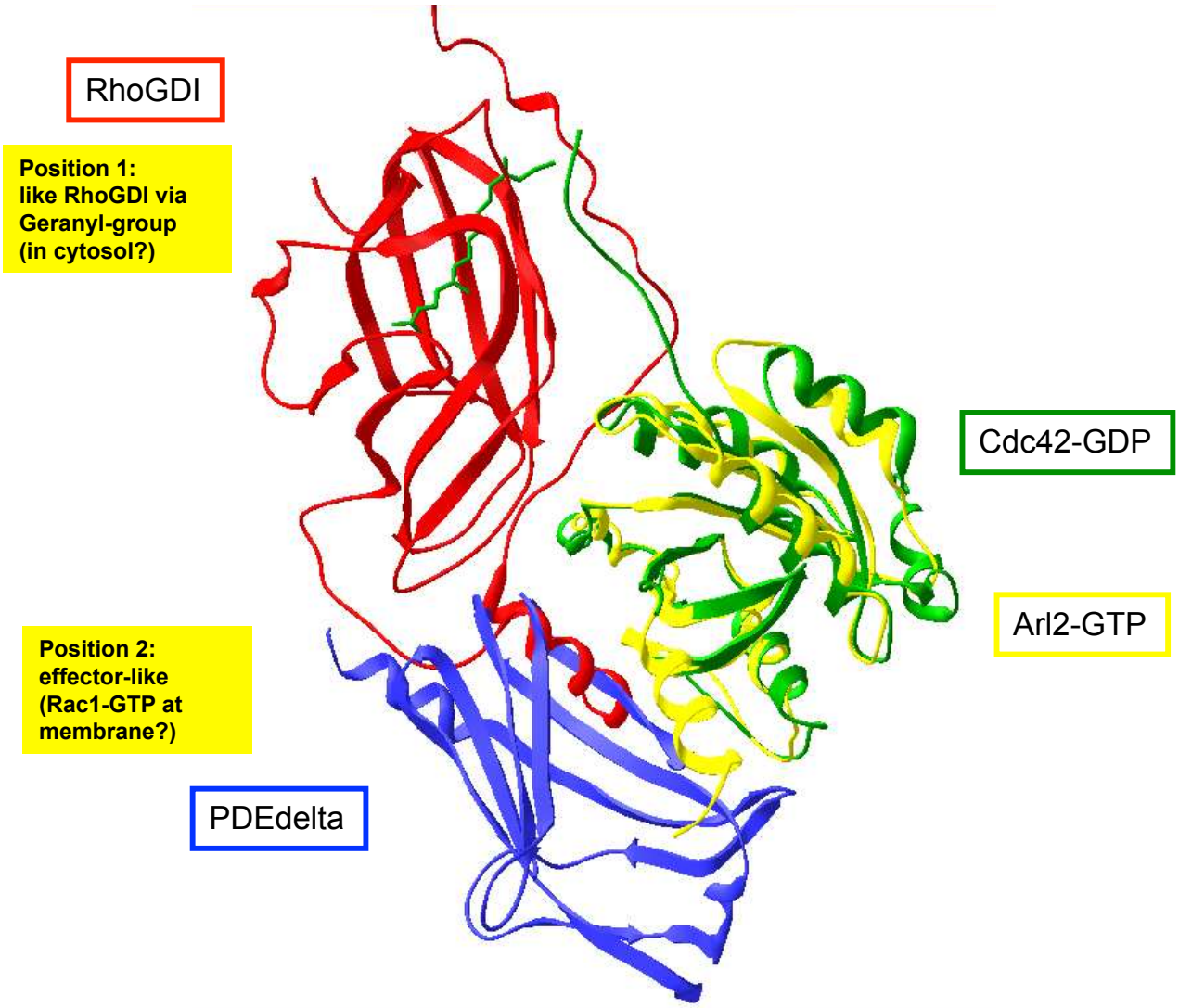
Supplementary Figure S4 Validation of protein interactions by immunostaining and affinity methods. (**A**) Colocalization of selected proteins was confirmed by immunoblot analyses of sucrose density gradient fractions. Antibodies used are indicated on the right; fraction number (from bottom to top) is indicated on the top of the panel. (**B**) Immunocoprecipitation with RhoA and Rac1 agarose-conjugated antibodies. Proteins from membranes (M) or cytosolic (soluble: S) fractions were preincubated with protein G-agarose and incubated with RhoA or Rac1 agarose-conjugated antibodies. Eluted proteins were resolved by SDS-PAGE and tested for the presence of Rhodopsin, RhoABC, Rac1, CRMP-2, ROCKII. (**C**) Concanavalin A pull-down of rhodopsin-associated proteins. Proteins from solubilized ROS membranes were incubated with Concanavalin A in the presence (+) or absence (-) of 0.2 mM AMM. Eluted proteins were resolved by SDS-PAGE and identified by immunoblotting, as indicated by the individual antibodies utilized on the right. (**D**) Immunoprecipitation (IP) of rhodopsin-associated proteins. Proteins from solubilized porcine ROS were immunoprecipitated with a rhodopsin-specific antibody (IP) or incubated with IgGs as a control (Cont). Eluted proteins were resolved by SDS-PAGE and identified by immunoblotting, as indicated by the individual antibodies utilized on the right. Additionally solubilized ROS lysate used for the IP was checked for b-tubulin as a control for using equal protein amounts.

Supplementary Figure S5



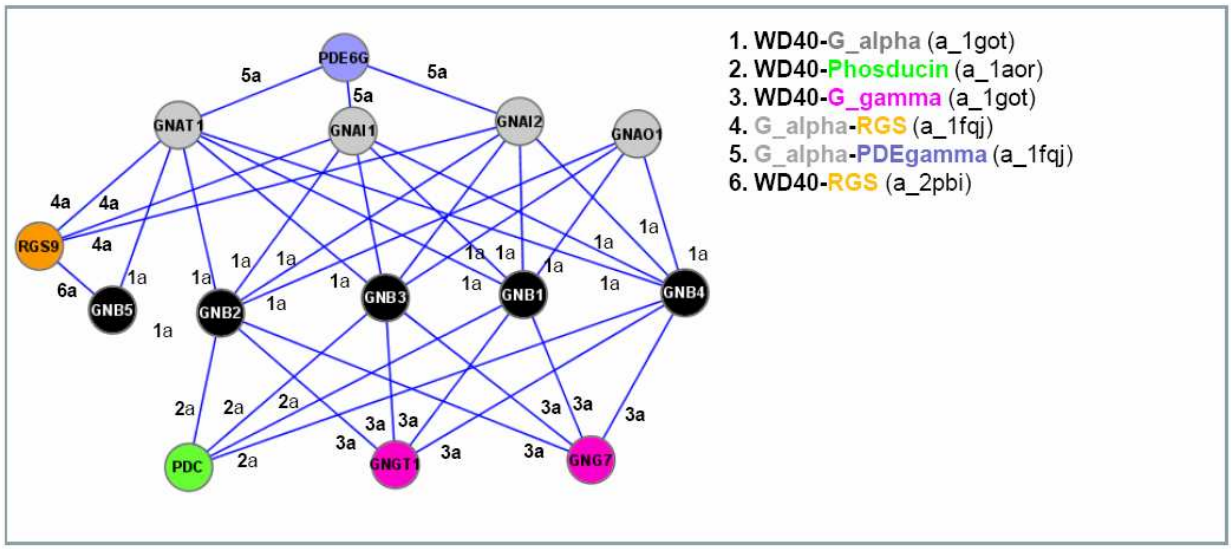
Supplementary Figure S5. Immunohistochemical analyses of porcine retina (controls). Cryostat sections of the retina were either stained with primary and secondary antibodies (red) against indicated proteins or only with secondary antibodies (mouse or rabbit) used throughout immunohistochemistry. Control sections omitting primary antibodies show no unspecific staining. Nuclei were counterstained (blue). For details please refer to Materials and methods.

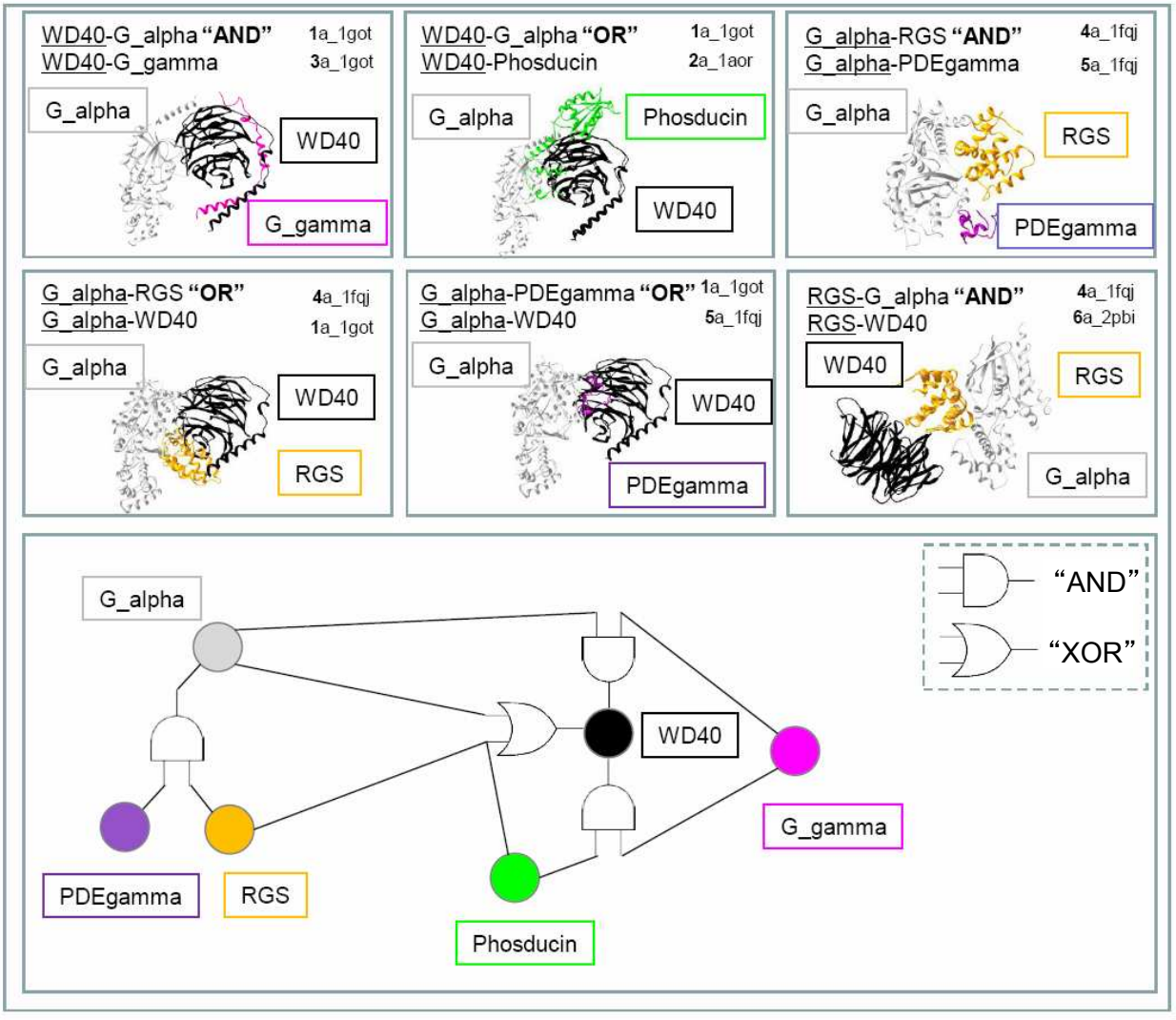
Supplementary Figure 6



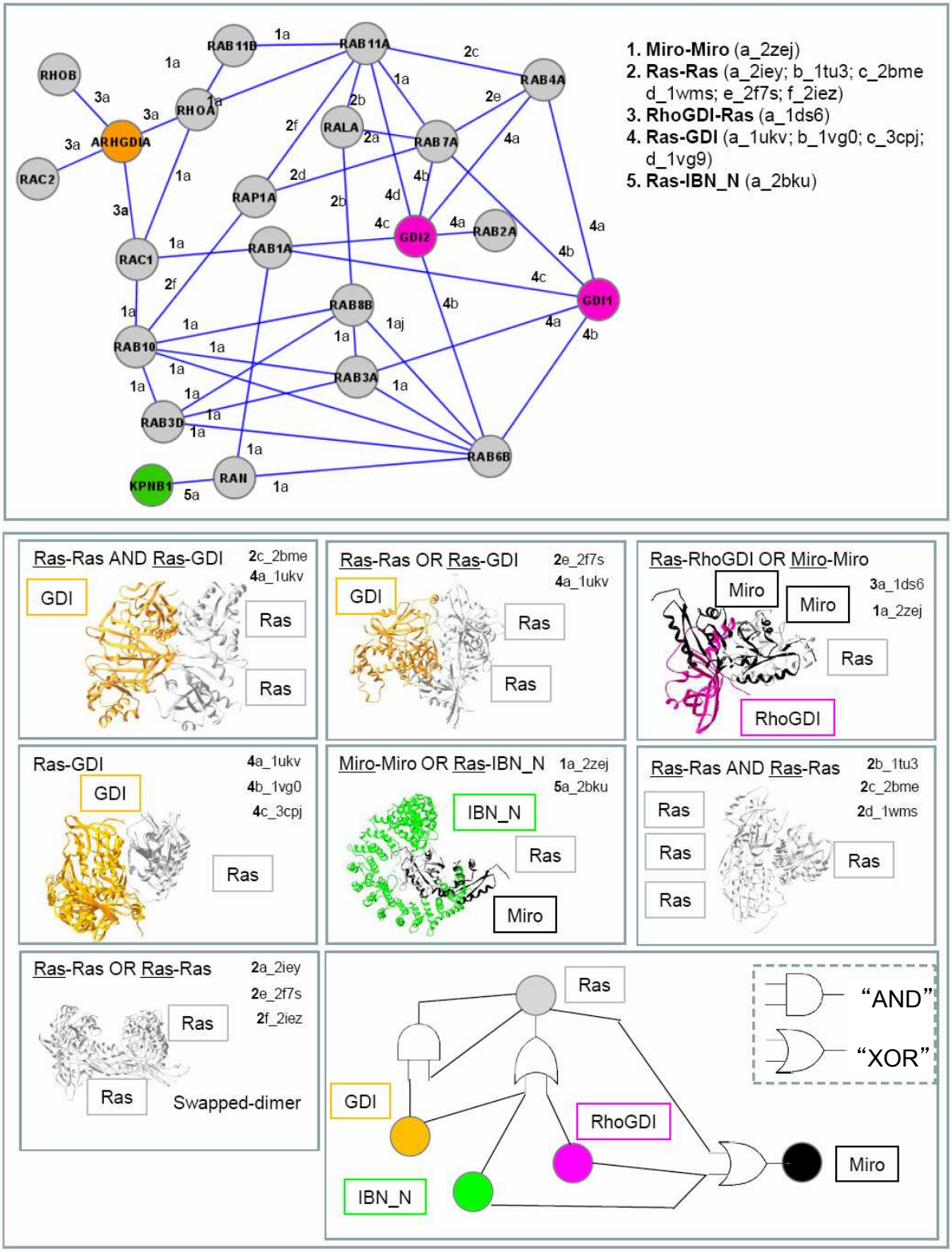
Supplementary Figure S6. Superimposition of the RhoGDI-Cdc42-GDP complex with Arl2-GTP/PDEdelta complex.

A

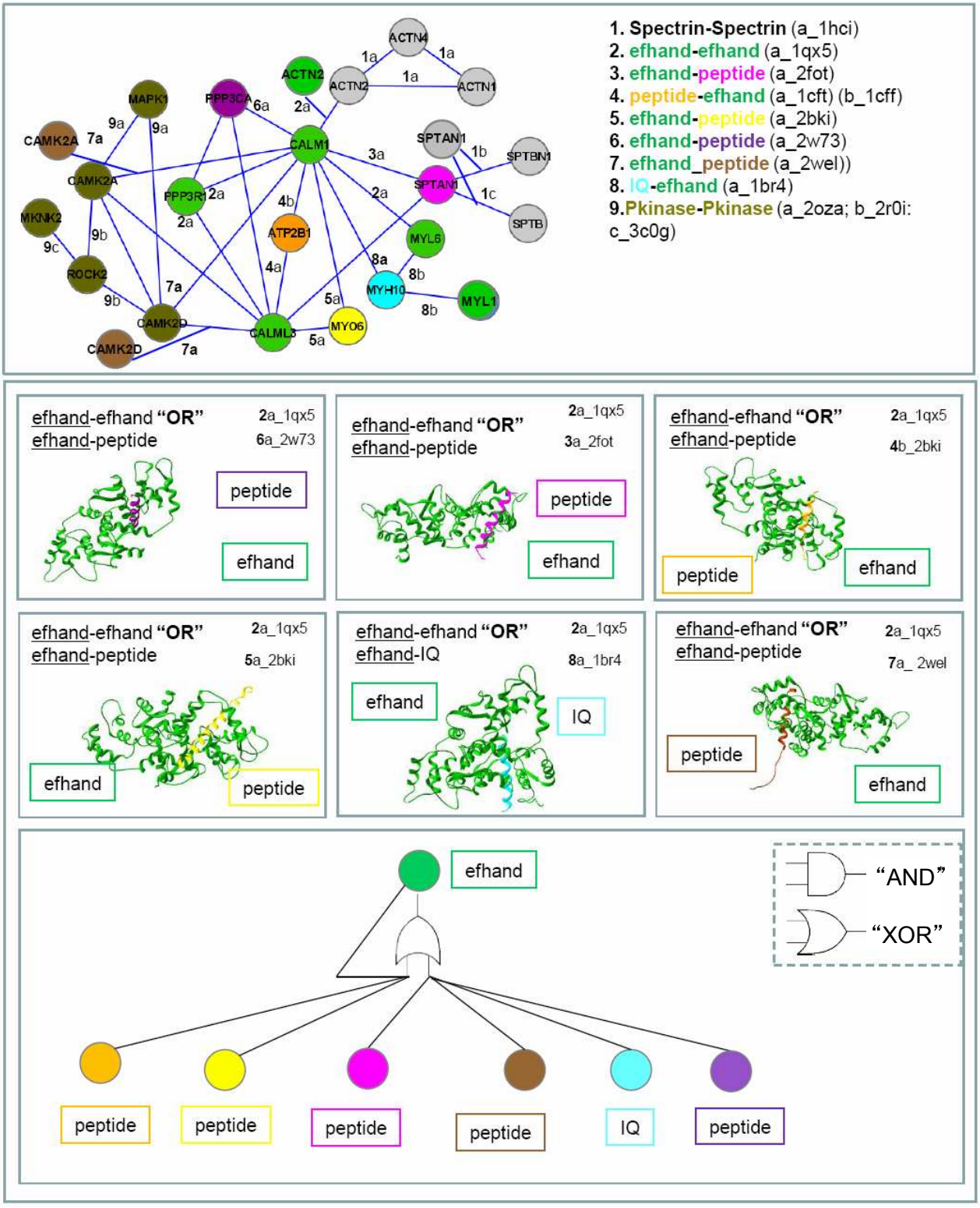

1. WD40-G_alpha (a_1got)
2. WD40-Phosducin (a_1aor)
3. WD40-G_gamma (a_1got)
4. G_alpha-RGS (a_1fqj)
5. G_alpha-PDEgamma (a_1fqj)
6. WD40-RGS (a_2pbi)
PDE6G
GNAT1
GNAI1
GNAI2
GNAO1
RGS9
GNB5
GNB2
GNB3
GNB1
GNB4
PDC
GNGT1
GNG7


WD40-G_alpha "AND" 1a_1got
WD40-G_gamma 3a_1got
G_alpha
WD40
G_gamma
WD40-G_alpha "OR" 1a_1got
WD40-Phosducin 2a_1aor
G_alpha
Phosducin
WD40
G_alpha-RGS "AND" 4a_1fqj
G_alpha-PDEgamma 5a_1fqj
G_alpha
RGS
PDEgamma
G_alpha-RGS "OR" 4a_1fqj
G_alpha-WD40 1a_1got
G_alpha
WD40
RGS
G_alpha-PDEgamma "OR" 1a_1got
G_alpha-WD40 5a_1fqj
G_alpha
WD40
PDEgamma
RGS-G_alpha "AND" 4a_1fqj
RGS-WD40 6a_2pbi
WD40
RGS
G_alpha
"AND"
"XOR"
G_alpha
WD40
G_gamma
PDEgamma
RGS
Phosducin

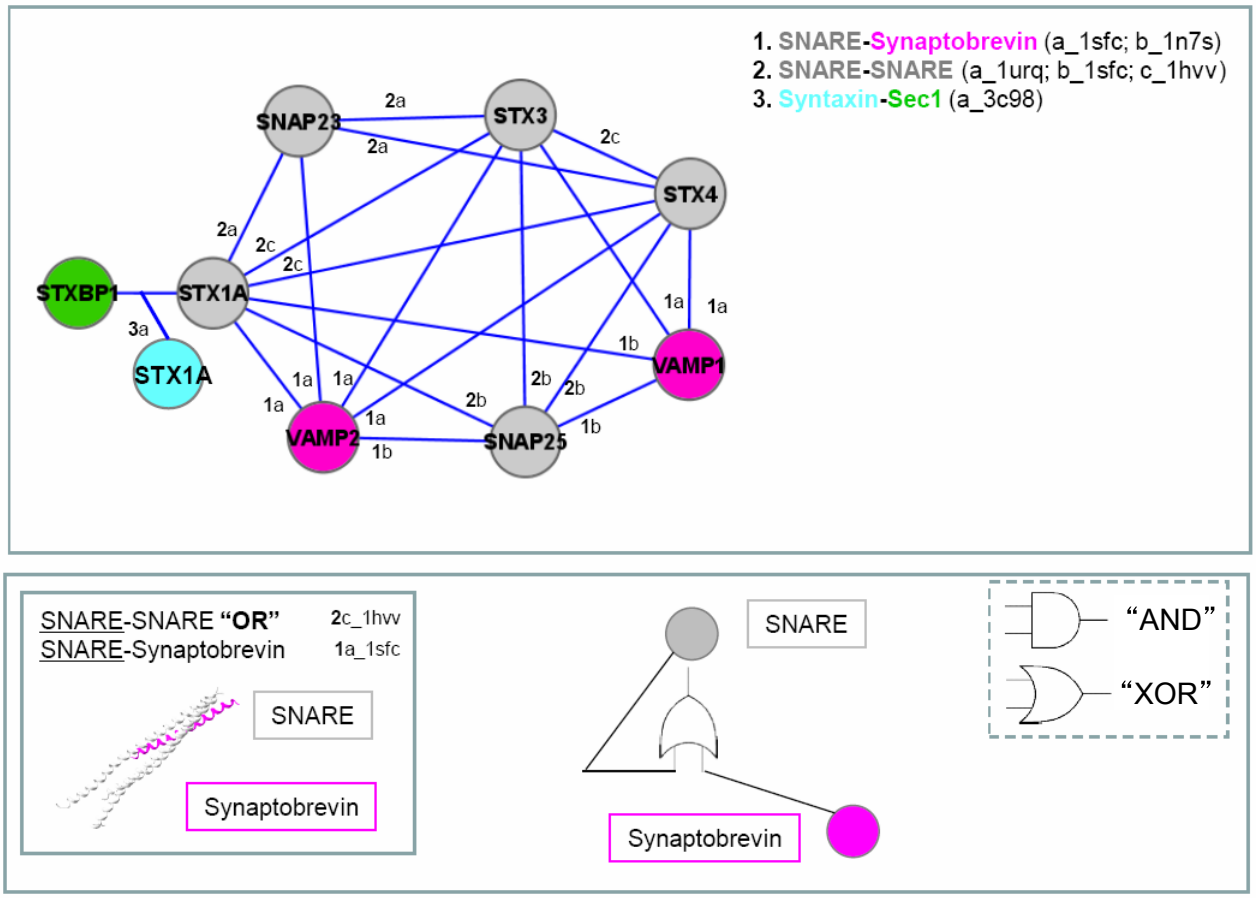
B


1. Miro-Miro (a_2zej)
2. Ras-Ras (a_2iey; b_1tu3; c_2bme d_1wms; e_2f7s; f_2iez)
3. RhoGDI-Ras (a_1ds6)
4. Ras-GDI (a_1ukv; b_1vg0; c_3cpj; d_1vg9)
5. Ras-IBN_N (a_2bku)
RAB11B
RAB11A
RHOB
RAB4A
RHOA
RALA
RAB7A
ARHGDIA
RAC2
RAP1A
GDI2
RAB2A
RAC1
RAB1A
GDI1
RAB8B
RAB10
RAB3A
RAB3D
RAB6B
KPNB1
RAN
Ras-Ras AND Ras-GDI
2c_2bme
4a_1ukv
GDI
Ras
Ras
Ras-Ras OR Ras-GDI
2e_2f7s
4a_1ukv
GDI
Ras
Ras
Ras-RhoGDI OR Miro-Miro
Miro
Miro
3a_1ds6
1a_2zej
Ras
RhoGDI
Ras-GDI
4a_1ukv
4b_1vg0
4c_3cpj
GDI
Ras
Miro-Miro OR Ras-IBN_N
1a_2zej
5a_2bku
IBN_N
Ras
Miro
Ras-Ras AND Ras-Ras
2b_1tu3
2c_2bme
2d_1wms
Ras
Ras
Ras
Ras
Ras-Ras OR Ras-Ras
2a_2iey
2e_2f7s
2f_2iez
Ras
Ras
Swapped-dimer
Ras
GDI
RhoGDI
IBN_N
Miro
"AND"
"XOR"

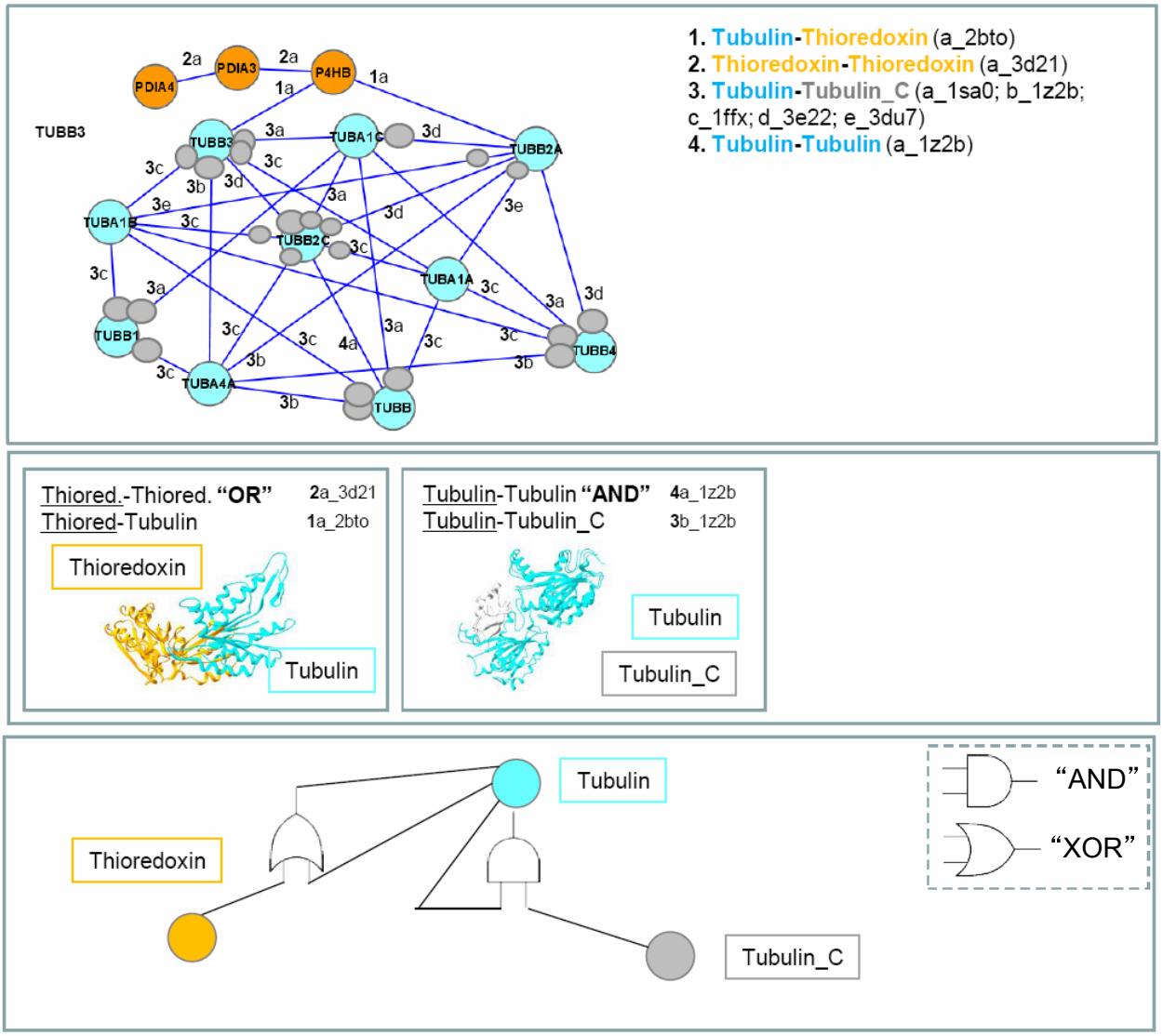
C


1. Spectrin-Spectrin (a_1hci)
2. efhand-efhand (a_1qx5)
3. efhand-peptide (a_2fot)
4. peptide-efhand (a_1cft) (b_1cff)
5. efhand-peptide (a_2bki)
6. efhand-peptide (a_2w73)
7. efhand_peptide (a_2wel))
8. IQ-efhand (a_1br4)
9.Pkinase-Pkinase (a_2oza; b_2r0i: c_3c0g)
ACTN4
ACTN2
ACTN1
MAPK1
PPP3CA
CAMK2A
CALM1
SPTAN1
SPTBN1
PPP3R1
ATP2B1
MYL6
SPTB
MKNK2
ROCK2
MYH10
MYL1
CAMK2D
CALML3
MYO6
efhand-efhand "OR"
efhand-peptide
2a_1qx5
6a_2w73
peptide
efhand
efhand-efhand "OR"
efhand-peptide
2a_1qx5
3a_2fot
peptide
efhand
efhand-efhand "OR"
efhand-peptide
2a_1qx5
4b_2bki
peptide
efhand
efhand-efhand "OR"
efhand-peptide
2a_1qx5
5a_2bki
efhand
peptide
efhand-efhand "OR"
efhand-IQ
2a_1qx5
8a_1br4
efhand
IQ
efhand-efhand "OR"
efhand-peptide
2a_1qx5
7a_ 2wel
peptide
efhand
efhand
"AND"
"XOR"
peptide
peptide
peptide
peptide
IQ
peptide

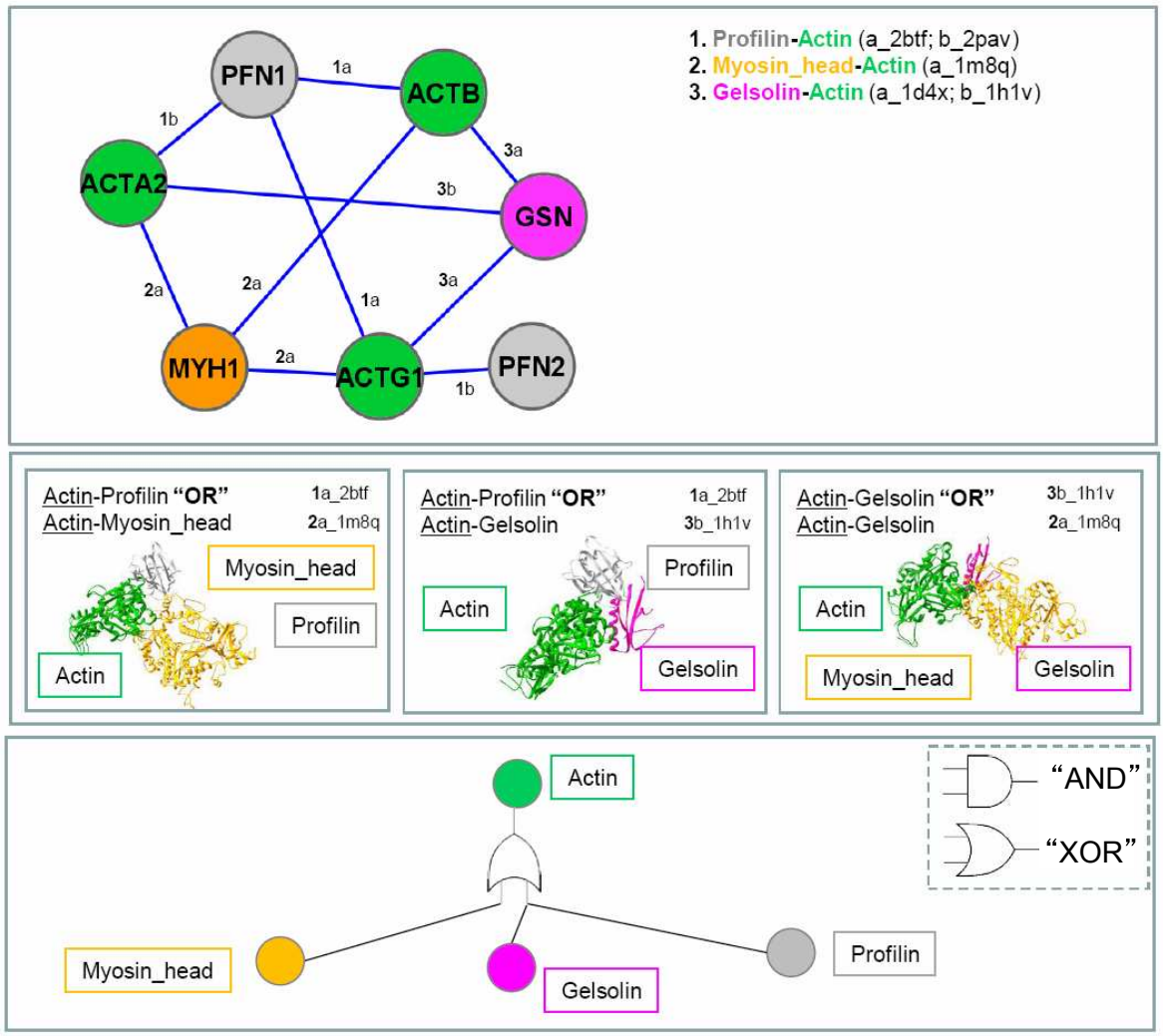
D


1. SNARE-Synaptobrevin (a_1sfc; b_1n7s)
2. SNARE-SNARE (a_1urq; b_1sfc; c_1hvv)
3. Syntaxin-Sec1 (a_3c98)
SNAP23
STX3
STX4
STXBP1
STX1A
STX1A
VAMP1
VAMP2
SNAP25
2a
2a
2c
2a
2c
2c
3a
1a
1a
1b
1a
1a
1a
2b
2b
2b
1a
1b
1b
SNARE-SNARE "OR" 2c_1hvv
SNARE-Synaptobrevin 1a_1sfc
SNARE
Synaptobrevin
SNARE
Synaptobrevin
"AND"
"XOR"

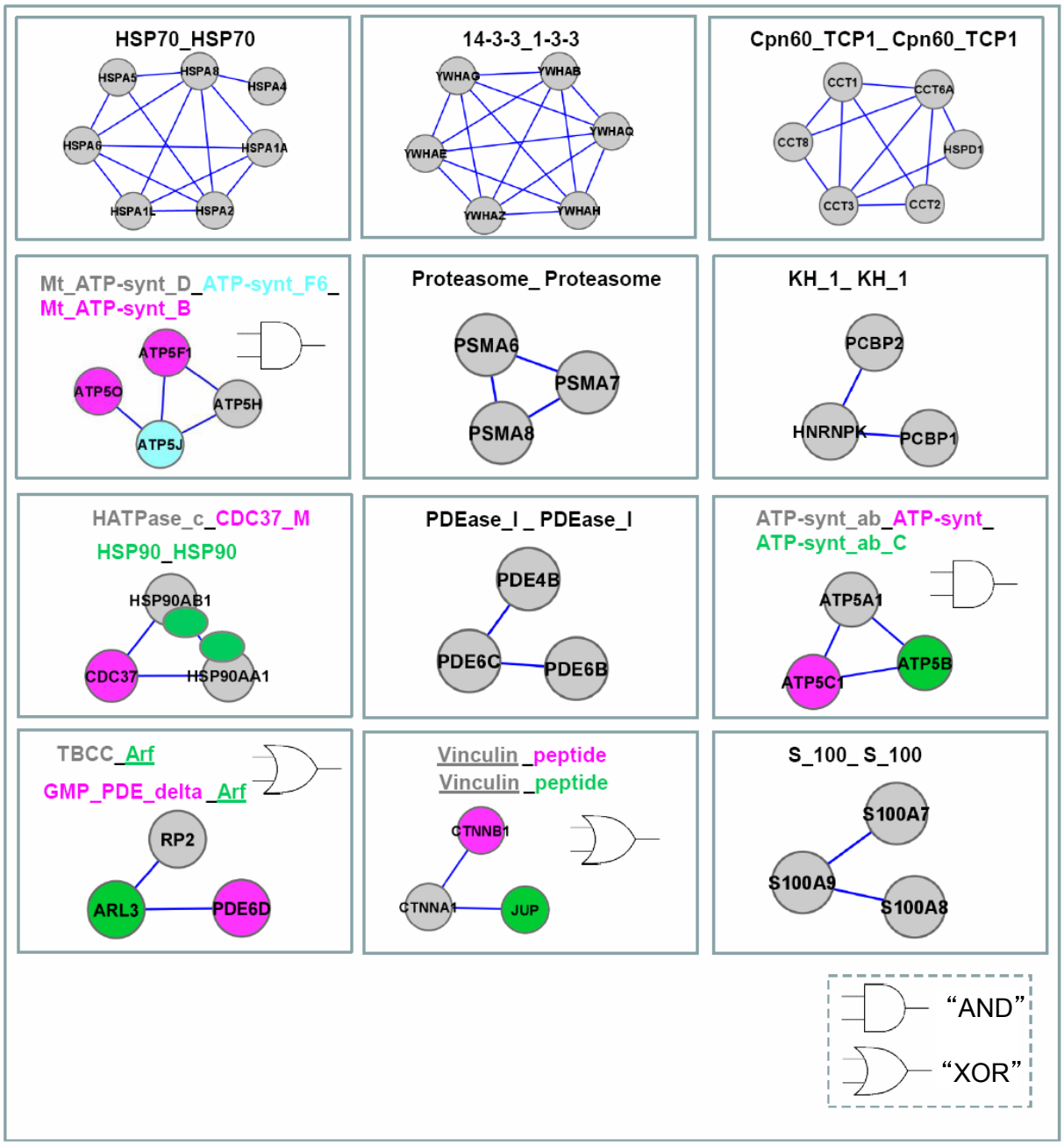
E


1. Tubulin-Thioredoxin (a_2bto)
2. Thioredoxin-Thioredoxin (a_3d21)
3. Tubulin-Tubulin_C (a_1sa0; b_1z2b; c_1ffx; d_3e22; e_3du7)
4. Tubulin-Tubulin (a_1z2b)
PDIA4
PDIA3
P4HB
TUBB3
TUBA1C
TUBB2A
TUBA1B
TUBB2C
TUBA1A
TUBB1
TUBA4A
TUBB
TUBB4
Thiored.-Thiored. "OR" 2a_3d21
Thiored-Tubulin 1a_2bto
Thioredoxin
Tubulin
Tubulin-Tubulin "AND" 4a_1z2b
Tubulin-Tubulin_C 3b_1z2b
Tubulin
Tubulin_C
Tubulin
Thioredoxin
Tubulin_C
"AND"
"XOR"

F

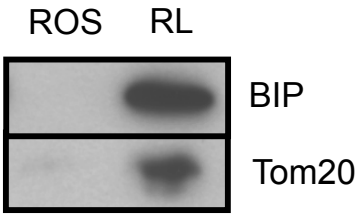

PFN1
ACTB
ACTA2
GSN
MYH1
ACTG1
PFN2
1a
1b
3a
3b
2a
2a
1a
3a
2a
1b
1. Profilin-Actin (a_2btf; b_2pav)
2. Myosin_head-Actin (a_1m8q)
3. Gelsolin-Actin (a_1d4x; b_1h1v)
Actin-Profilin "OR" 1a_2btf
Actin-Myosin_head 2a_1m8q
Myosin_head
Profilin
Actin
Actin-Profilin "OR" 1a_2btf
Actin-Gelsolin 3b_1h1v
Profilin
Actin
Gelsolin
Actin-Gelsolin "OR" 3b_1h1v
Actin-Gelsolin 2a_1m8q
Actin
Myosin_head
Gelsolin
Actin
"AND"
"XOR"
Myosin_head
Gelsolin
Profilin

G



Supplementary Figure S7. Superimposition of complex structures. Interacting domains from complex structures were extracted and superimpositions were done, if similar domains are involved in more than one interaction. Superimpositions were analyzed manually using the SwissPdb Viewer Software, and if two partner domains for the same domain were non overlapping, assigned as compatible (“AND’), and if overlapping, assigned as mutually exclusive (“XOR”).

Supplementary Figure S8



Supplementary Figure S8 Analysis of ROS preparation purity. Porcine ROS and retina lysate (RL) (20µg each) were resolved by SDS-PAGE, immunoblotted and tested for the presence of specific endoplasmatic reticulum (BIP) and mitochondrial (Tom20) proteins, generally only found in inner segments (RIS).

3.2 SAPIN: Structural Analysis of Protein Interaction Networks

Proteins are involved in all the biological processes taking place inside the cell. They can assemble to form stable and large molecular machines or bind to each other in a more dynamical way. The high-throughput identification of protein interactions initiated a decade ago has produced an unprecedented large amount of data, usually represented by networks, where proteins are shown as nodes and the interactions between them as edges. Although these networks have helped to derive important characteristics inherent to biological systems, they do not represent the information on competition between proteins binding to a similar central 'hub' node. Since many proteins have many more interaction partners than surface available for binding, it is obvious that no all interacting proteins can bind at the same time, and that there will be competition among them. Thus, the knowledge of protein interfaces contained in 3D structures can add this missing information into protein interaction networks. However, the structural coverage at the interactome level remains low (there is a structure only for less than 10% of protein interactions in human), due to the current limitations of protein structure determination methods.

To overcome this limitation, many computational approaches, like comparative modeling and docking methods, have been developed with the objective of extending the knowledge of protein interfaces to a larger amount and to unravel the molecular mechanisms of protein interactions. However, these approaches can be computationally expensive and time consuming and may consequently be difficult to apply on large data sets. In addition, they often deal with pairs of proteins or domains, without taking into account the context in which they occur at the network level.

Here, we implemented an automated method, SAPIN, to analyze protein interactions from a structural point of view. Our approach is based on the modular property of proteins organized in domains and the observation that similar domain pairs usually interact in a similar way (Aloy *et al.*, 2003). Given a set of protein interactions and its related sequences, SAPIN first predicts structural interactions, using structures of domain-domain and domain-peptide interactions. Using 3D structures of domain interactions is valuable since it has recently been observed that they can increase

the structural coverage of around 50 % for the human interactome (Stein *et al.*, 2011c). However, the classification of domain interaction interfaces has shown that in many cases, a given pair of domain interaction could show multiple possible orientations (Kim *et al.*, 2006b). So the knowledge of protein domains is not enough in these cases and the potential templates need to be further analyzed in order to select the most suitable one. To cope with this issue, SAPIN embeds InterPreTS (Aloy & Russell, 2002), which evaluates an interface based on empirical potential of the interacting residues. Once the structural interactions have been predicted within the network, SAPIN identifies in a third step protein interactions that are compatible and mutually exclusive. This is particularly relevant, as it provides an important missing feature to protein interaction networks: competition for binding. We developed SAPIN as fully automated procedure, which has been tested on a large number of cases. Based on these results, we were able to predict correctly the compatible interactions tested in 99% of the cases at a reasonably low threshold of clashing residues at the interface (i.e. 15%).

Finally we made SAPIN available through a web server to be shortly accessible through the URL: http://sapin.crg.es. The following manuscript is ready for submission to *Bioinformatics*, as an application note.

SAPIN: Structural Analysis for Protein Interaction Networks

Anne Campagna[1], Peter Vanhee[1], Luis Serrano[1,2*] and Christina Kiel[1*]

[1] MBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), UPF, Barcelona, Spain

[2]Institució Catalana de Recerca I Estudis Avançats (ICREA)

*To whom correspondence should be addressed:

christina.kiel@crg.eu; luis.serrano@crg.eu

Running title: SAPIN webserver

Category: Structural bioinformatics

ABSTRACT

Summary: Protein interaction networks are widely used to depict the relationships between proteins within a cell or a sub-cellular compartment as nodes connected with edges. These networks often lack dynamic information and do not inform if there is incompatibility between binding partners. The three-dimensional (3D) structures of proteins remain the most valuable source for understanding the molecular details of binding. Here we present a web server, SAPIN, which is dedicated to the structural analysis of protein interactions (between globular domains or domain-peptide interactions), ranging from a single pair to larger networks. It first identifies the parts of the proteins that could be involved in the interaction (i.e. domains or linear motifs) and in a second step brings these data together with available structural information to provide a template for the interaction. Finally, the algorithm performs the analysis of compatible and exclusive interactions among the previously structurally characterized interactions, adding an important missing feature into classical networks: the competition for interacting partners. Finally, the results are displayed in Cytoscape Web.

Availability: The SAPIN server is available at URL http://sapin.crg.es.

Contact: anne.campagna@crg.eu; christina.kiel@crg.eu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In the past decade, an important effort has been made to identify all proteins and their interactions within many species. It is now clear that the majority of the proteins exercise their functionality not in isolation but as part of protein complexes with stable or transient interactions (Nooren and Thornton, 2003). Thus both, experimental and computational methods have become of increasing importance to understand which proteins are involved in a biological process and, how they interact with each other to form complexes and achieve their function. Experimental methods, applied in a high-throughput fashion, have produced large catalogues of interactions (Gavin *et al.*, 2002; Ho *et al.*, 2002; Rual *et al.*, 2005). Whereas such networks are convenient to provide a global view of the content of a given cell or sub-cellular compartment, they give a rather static picture of complex behaviors and highly dynamical events that occur within the cell. The final result is that one protein could have almost one partner per amino acid, i.e. p53 is reported to have 380 partners, according to the STRING database (Szklarczyk *et al.*, 2011) at the highest confidence, while it is made of 392 amino acids; it is obvious that not all interaction partners can bind at the same time. In addition, part of the data deposited in these databases do not come from binary interaction detection methods but from co-immunoprecipitation (co-IP) or TAP-TAG methods (reviewed in Drewes and Bouwmeester 2003), which provide purified fractions of proteins associated with a bait protein. At the moment, only statistical approximations, such as the *socioaffinity index* (Gavin *et al.*, 2006) are used to decide which of the putative interactions detected in a co-IP or TAP-TAG experiments are direct.

These two problems could be partly solved by using 3D structures of proteins and protein complexes. 3D information can determine competing interactions and combined with protein interaction data, it can support weak pieces of evidence for binary interactions. Moreover, based on the observation that homologous pairs of binding proteins tend to use the same interaction topology (Aloy *et al*. 2003), the structural knowledge of protein interactions can be extended to the interactions whose structure has not been resolved and thus contribute to filling the gap between the low number of structures and the large collection of sequences available.

Here we introduce SAPIN, a framework dedicated to the Structural Analysis of Protein Interaction Networks. It encompasses many features allowing (i) a full analysis of the protein sequence for the identification of the parts potentially involved in an interaction, (ii) a mapping

of the available structural data involving the previously identified parts, and (iii) the identification of compatible and mutually exclusive interactions at the network level.

2 SAPIN WORKFLOW

SAPIN has been implemented using Python programming language. Figure 1 provides an overview of the method, which consists of three main parts:

a. Sequence analysis and sequence-based interactions

The objective of this part is to predict the domains and linear motifs of a protein and which could mediate the interaction with binding partners. The applications used here require the protein sequences as input and return as output: (i) the predicted domain composition using HMM collection from PFAM (Finn *et al.*, 2010)); (ii) the possible phosphorylation sites and motifs that potentially can bind to globular domains, derived from experimental data (Phospho.ELM; Dinkel *et al.*, 2011) or from prediction methods (Scansite; Obenauer, 2003) and NetPhorest; Miller *et al.*, 2008); (iii) predictions that identify disordered regions (Disopred; Ward *et al.*, 2004); and (iv) secondary structure element composition complete sequence analyses (Jones, 1999; Bryson *et al.*, 2005). Identifying disordered regions in proteins is particularly relevant in combination with predicted phosphorylation sites or binding motifs, as these are usually located in unstructured parts of the protein (Dyson and Wright, 2005).

b. Search for structural templates of domain-domain or domain-motifs interactions

Here, we search for structural information for an interaction, which could be an original x-ray structure or we provide a model. First, the Protein Data Bank (Rose et al., 2011) is searched based on sequence similarity (using BLAST; Camacho *et al.*, 2009). For a given pair of proteins, both sequences are searched independently. If the search identifies a pair of homologous proteins in interaction within a PDB entry, the corresponding domain interaction is then derived using the domain composition of the proteins. Structural templates for the interaction are selected when the sequence identity for both proteins is above 70 %, which is a reliable threshold for interaction architectures according to Aloy *et al.*, (Aloy *et al.*, 2003). If the search against the PDB does not provide any result, the 3DID database (Stein et *al.*, 2005; Stein *et al.*, 2009) (filtered for crystal packing artifacts, see Supplementary data) is further searched for a potential structure of domain-domain or domain–motif interactions that could model the interaction. The resulting matches are

then evaluated using InterPreTS (Aloy and Russell, 2002; Aloy and Russell, 2003) in order to select the most relevant domain interaction.

c. Identification of compatible and mutually exclusive complexes.

Finally, the structural information is used to identify the compatible and mutually exclusive interactions within the network. At this level, we are looking at each protein and its structurally identified interacting partners. Thus, a protein needs to have at least two partners for which there are structural data to determine whether their binding is compatible or not with the protein of interest. To this end, all the domain-domain and/or domain-motif interaction structures are structurally aligned on their reference domain in the input protein using the pairwise structural comparison tool DaliLite (Holm and Park, 2000). The interacting domains are then analyzed pairwise for backbone van der Waals clashes (above 1kcal/mol) using the empirical force field FoldX (Schymkowitz *et al.* 2005). If less than 15% of the residues involved in the interface have clashes, the interactions are compatible, and two interfaces are assigned to the protein (see Supplementary data; and see website for more detailed information of how to interpret the results). If there are more than 15% of residues with clashes at the interface, the interactions are mutually exclusive, and one interface is assigned to the protein. All the interacting proteins are analyzed in this iterative process, and assigned either to an existing interface or to a new one, if they are found to be compatible with any of the analyzed interacting domain. At the end of the process, each protein is defined by its domains (or linear motifs) and each of these domains has one or several interfaces, depending on the interacting proteins.

3 THE SAPIN WEBSERVER

SAPIN is accessible through a web portal at http://sapin.crg.es. The portal is built on the open – source Drupal Content Management System for full flexibility. The required input is an "interaction file", in tab-separated two-column format, where each line refers to a pair of proteins and a fasta-formatted file containing the protein sequences from the whole dataset that has to be analyzed. The identifier line in the fasta file needs to correspond to the interaction file. The pipeline is designed to deal preferably with UNIPROT (The UniProt Consortium 2011; Jain *et al.* 2009) accession numbers, as all the resources used in the pipeline have been converted into UNIPROT accession numbers to facilitate the process.

For each pipeline step described above, we provide a user-friendly interface to visualize the results. The sequence features for each protein can be browsed by simply clicking on a residue to display the detailed results from the different predictions. The structural analysis can be viewed through different tables, which summarize the output of the search against the PDB, the mapping from the 3DID database and the corresponding InterPreTS scores and the identified domain-motif interactions. Further, the structural analysis identifying compatible and mutually exclusive interactions is showed in a table and in an interactive network browser (Lopes *et al.* 2010). Finally, we represent the structural information by adding extra nodes: in addition to the classical nodes representing the proteins, we show the domain with an extra node and the different interfaces to which the interacting partners are binding with another node linked to the domain.

4 DISCUSSION AND CONCLUSION

Protein interaction detection methods applied in a high-throughput way have provided large catalogues of valuable data. However, these data contain false positives and it is not straightforward to extract binary interactions from co-purified protein complexes. SAPIN is a framework that brings together protein interaction networks and structural data, with the objective of reaching a better understanding of how proteins interact with each other. It predicts structural interactions based on sequence data from proteins that have been experimentally co-purified within a complex. It then uses the knowledge of interaction interfaces to identify compatible and exclusive interactions. However, this approach is based on single domain-domain interactions, without taking into account the conformational arrangement of the full-length proteins. This method highlights the principle of competition, which could be important in signal transduction pathways. Further, it could be combined with statistical approaches (e.g. socio affinity index) in order to describe more accurately the organization of protein complexes.

Funding: This work was supported by the EU (PROSPECTS, grant agreement number HEALTH-F4-2008-201648 to Luis Serrano).

REFERENCES

Aloy,P. and Russell,R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics (Oxford, England)*, 19, 161-2.

Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 5896-901.

Aloy,P. et al. (2003) The Relationship Between Sequence and Interaction Divergence in Proteins. *Journal of Molecular Biology*, 332, 989-998.

Bryson,K. et al. (2005) Protein structure prediction servers at University College London. *Nucleic acids research*, 33, W36-8.

Camacho,C. et al. (2009) BLAST+: architecture and applications. *BMC bioinformatics*, 10, 421.

Dinkel,H. et al. (2011) Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic acids research*, 39, D261-7.

Drewes,G. and Bouwmeester,T. (2003) Global approaches to protein – protein interactions. *Current Opinion in Cell Biology*, 199-205.

Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nature reviews. Molecular cell biology*, 6, 197-208.

Finn,R.D. et al. (2010) The Pfam protein families database. *Nucleic acids research*, 38, D211-22.

Gavin,A.-C. et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440, 631-6.

Gavin,A.-C. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415, 141-7.

Ho,Y. et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415, 180-3.

Holm,L and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics (Oxford, England)*, 16, 566-7.

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

Jones,D T (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292, 195-202.

Lopes,C.T. et al. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics (Oxford, England)*, 26, 2347-8.

Miller,M.L. et al. (2008) Linear motif atlas for phosphorylation-dependent signaling. *Science signaling*, 1, ra2.

Nooren IM, and Thornton JM. (2003) Diversity of protein-protein interactions. *EMBO J*, 22:3486-92.

Obenauer,J.C. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, 31, 3635-3641.

Rose,P.W. et al. (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research*, 39, D392-401.

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

Schymkowitz,J.W.H. et al. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 10147-52.

Snel,B. et al. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, 28, 3442-4.

Stein,A. et al. (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic acids research*, 37, D300-4.

Stein,A. et al. (2005) 3Did: Interacting Protein Domains of Known Three-Dimensional Structure. *Nucleic acids research*, 33, D413-7.

Szklarczyk,D. et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39, D561-8.

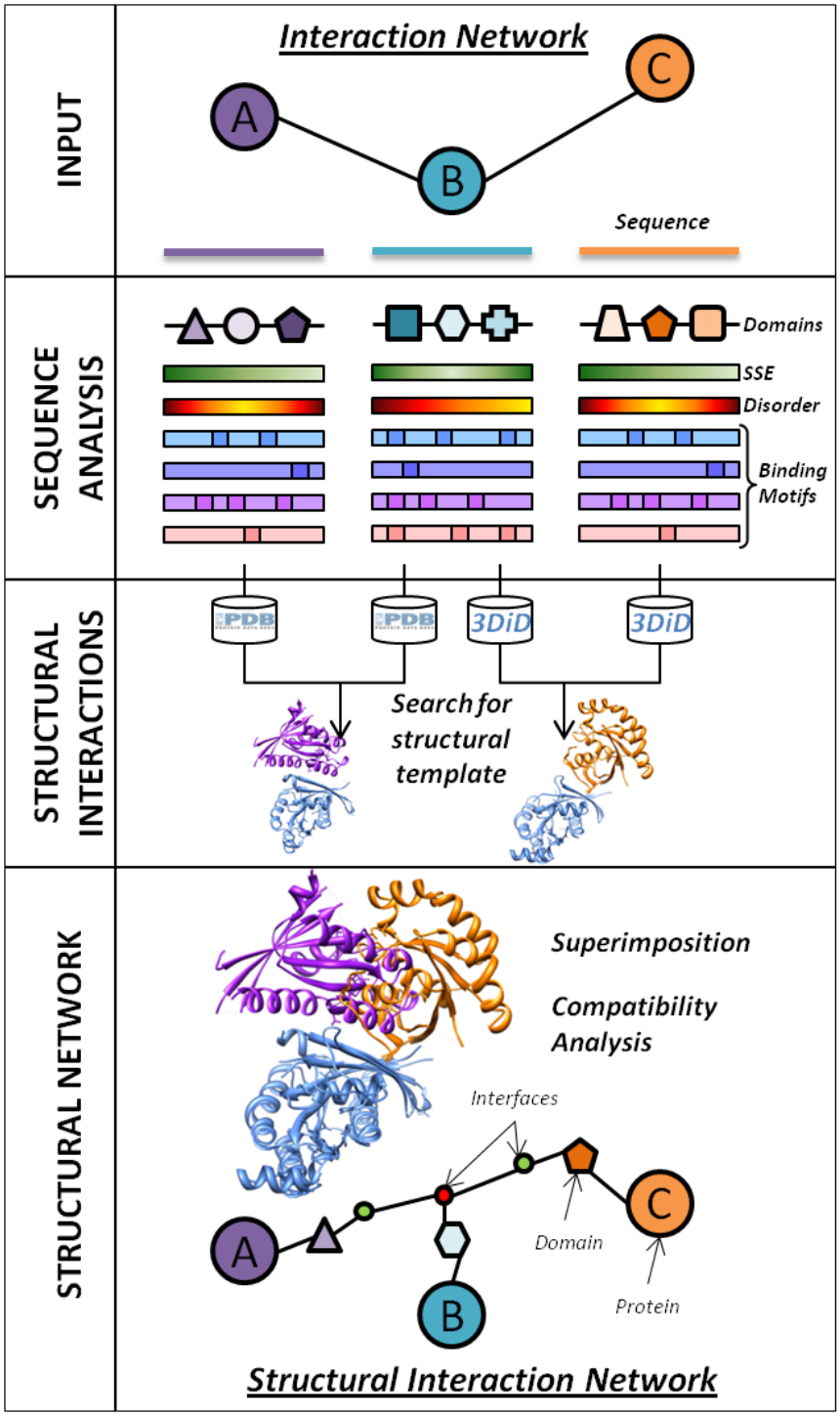
The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research*, 39, D214-9.

Ward,J.J. et al. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics (Oxford, England)*, 20, 2138-9.

FIGURE LEGEND

Fig. 1. Overview of the SAPIN webserver. As an input, the pipeline takes a network, in the example shown, with three proteins (A, B, C), and two interactions (A-B, A-C), and the related protein sequences. First, the sequences are analyzed to determine the domain composition, secondary structure elements (SSEs), disordered regions, binding motifs and phosphorylation sites. Then, the available structural data is mapped to identify a potential structural template for each interaction. Finally, if a protein has at least two structural interacting partners, the interactions are superimposed on the reference domain (in this case the domain from protein A, in blue) and the interacting domains are analyzed for clashing. This is showed in a final structural interaction network by adding nodes for the domains involved in the interactions, and for the interfaces through which the binding takes place.

Figure 1 (Campagna et al.)



SAPIN: Structural Analysis for Protein Interaction Networks

Supplementary material and methods

Filtering of 3DiD database

The database of 3D Interacting Domains (3DID) (Stein et al. 2011) is a collection of domain-domain and domain-peptide interactions in proteins for which high-resolution three-dimensional structures are known. It results from the Protein Data Bank being mapped with PFAM domains, and the corresponding structures have been extracted and organized. We have observed that 3DID database contains crystal-packing contacts, especially regarding the interactions of homodimers. In order to improve the quality of the results provided by our method, we filtered out the structures coming from 3DiD containing crystal-packing contacts. From the last major release of the PDB (Rose et al. 2011), the authors of the resolved structures are now providing the information on the biological units contained in the PDB entries. We parse all the PDB file headers, extract the information of biological interactions and use it to analyse 3DiD. The resulting filtering made the total number of structures decrease of 40% (Figure S1) and around 20% of the domain interactions from 3DID are not present in our filtered version. In the cases where the author annotation was not available, we used the Protein Quaternary Structure server (PQS) (Henrick 1998).

Estimation of the superimposition error

In order to evaluate the threshold of clashing residues for which we decide if the interactions are compatible or mutually exclusive, we designed the following experiment (see Figure S2 for an overview): we selected complexes involving three domains belonging to three different polypeptide chains. We set one the domains as

a reference domain, and split the complex into two binary domain interactions, both containing the reference domain and one of the two other domains (Figure S2-A). We have thus two pairs of interacting domains that are compatible and we want to see if superimposing them on homologous domains can affect relative position to each other, and generate clashes that would make them incompatible. A first round of superimpositions is done by structurally aligning independently the two domain interactions on a set of structures of domains homologous to the reference domain (Holm and Park 2000) (figure S2-B, part 1). We then substitute the reference domain by each one of the homologues (figure S2-B, part 2) and perform all the pair wise superimpositions of the hybrid domain interactions (figure S2-B, part 3). The interface of the domains interacting with the reference domain if then analyzed using the empirical force field FoldX (Schymkowitz et al. 2005) (figure S2-B, part 4). In particular, the Van der Waals clashes (above 1kcal/mol) involving the backbone of the residues located at the interface of the domain interaction are evaluated. Figure S3 shows the results for five tests carried out on five complexes (see details in tables S1). For one of these complexes (Complex1), all the different combinations result in no clashes at all. For Complex4 and Complex5, most of the combinations show no clashes, as the median and quartiles are very close to zero on the boxplot representation (figure S2-A). For Complex2 and Complex3, the distributions exhibit a similar trend, with most of the combinations having no clashes. The distributions in these cases are more spread but the majority of the cases do not exhibit more than 15% of residue clashing in the interface. Finally, taken all together, the results show that 2.7% of the complexes are analyzed as not compatible at a threshold of 1% of clashing interface residues (Figure S4). According to these results, we set the threshold for clashing residues at 15% since at this value, 99% of the reconstructed complexes are correctly evaluated.

REFERENCES

Henrick,K. (1998) PQS: a protein quaternary structure file server. *Trends in Biochemical Sciences*, 23, 358-361.

Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics (Oxford, England)*, 16, 566-7.

Rose,P.W. et al. (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research*, 39, D392-401.

Schymkowitz,J.W.H. et al. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 10147-52.

Stein,A. et al. (2011) 3Did: Identification and Classification of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucleic acids research*, 39, D718-23.

Figure S1

Representation of crystal packing in 3DiD database. The biological interfaces are shown in blue (as provided by the author annotation from the PDB entries) and dark orange (as provided by PQS). The interactions considered as crystal packing are shown in light orange.

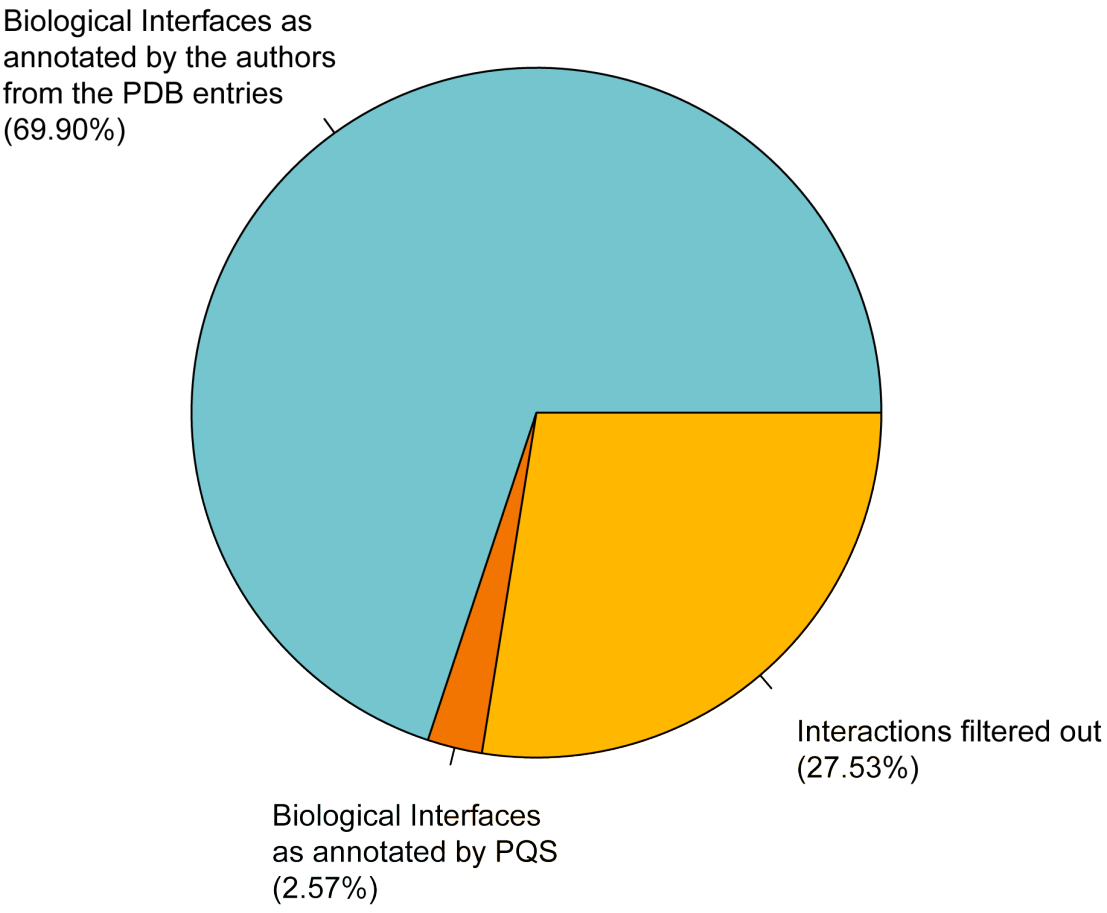


Table S1

Details of the complexes used to analyze the superimposition error. Each of the complex consists of a trimer with three domains in interaction belonging to three different polypeptide chain from a PDB entry.

Complex	PDB Id	Reference Domain				Interacting domains			
		Domain Name	Chain	Residue Start	Residue End	Domain Name	Chain	Residue Start	Residue End
1	1A3F	Phospholip_A2_1	A	1	118	Phospholip_A2_1	B	1	118
						Phospholip_A2_1	C	1	118
2	1JSU	Cyclin_N	B	181	307	CDI	C	30	80
						Pkinase	A	13	286
3	1JSU	Pkinase	A	13	286	Cyclin_N	B	181	307
						CDI	C	30	80
4	1WEJ	Cytochrom_C	F	3	102	V-set	H	1	115
						V-set	L	1	106
5	1WEJ	V-set	H	1	115	V-set	L	1	106
						Cytochrom_C	F	3	102

Figure S2

Workflow for the analysis of the superimposition error. A. One domain from the complex is set as the “reference domain” (grey). The complexes are then split into two domain interactions, each of them containing the reference domain (grey) and the interaction domains (orange and blue). B. Each domain-domain interaction (ddi1 and ddi2) is superimposed using their reference domain onto a selection of structures of homologous domains (1). Each one of these homologous domains is then used to build a hybrid domain-domain interaction (2)(ddi1bis and ddi2bis). Theses two domain interactions are then superimposed on each other, based on their substituted reference domain (3). Finally, the interface of the interacting domains is analyzed for backbone clashes for residues located at the interface (4).

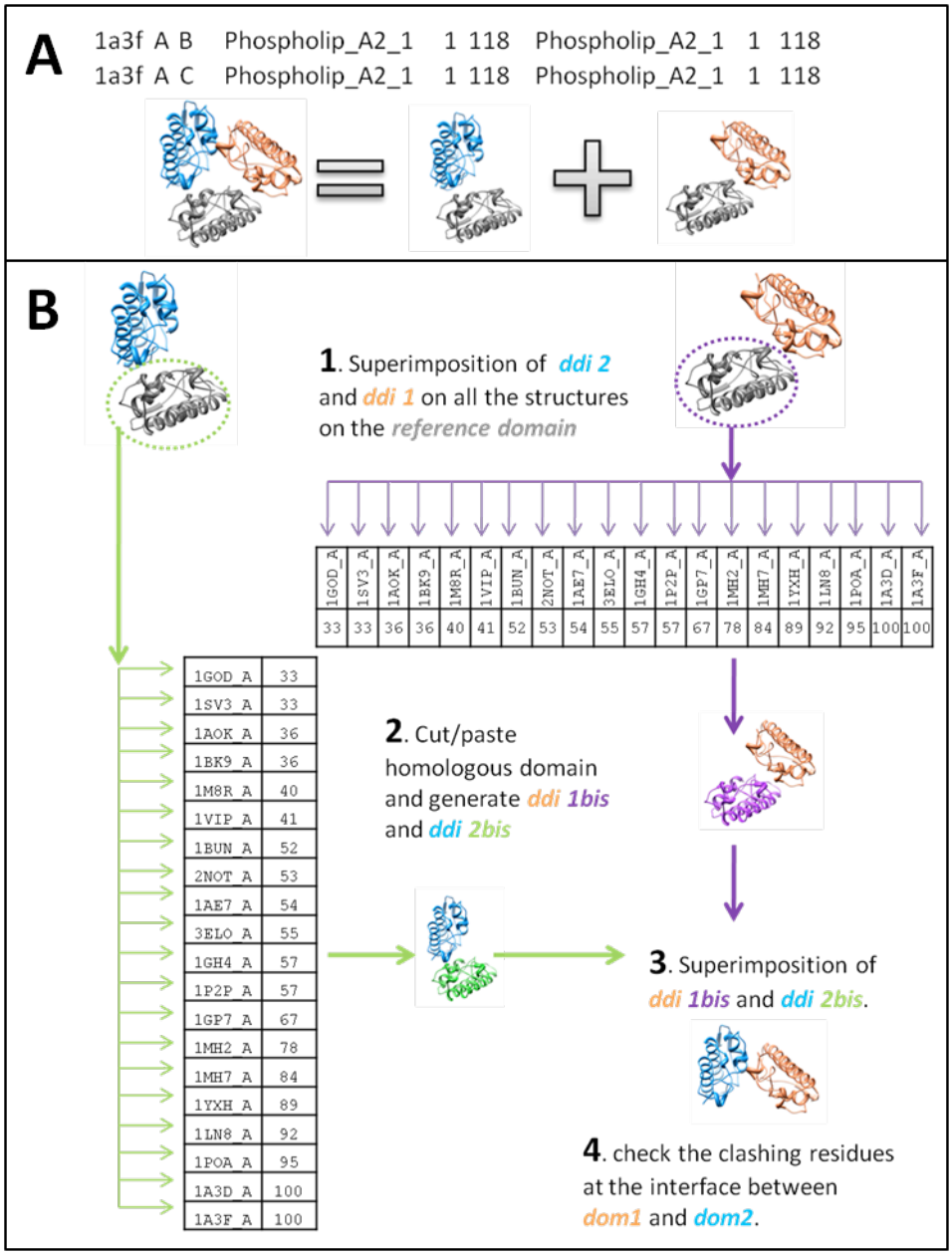

A
1a3f A B Phospholip_A2_1 1 118 Phospholip_A2_1 1 118
1a3f A C Phospholip_A2_1 1 118 Phospholip_A2_1 1 118
B
1. Superimposition of ddi 2 and ddi 1 on all the structures on the reference domain
1GOD_A 1SV3_A 1AOK_A 1BK9_A 1M8R_A 1VIP_A 1BUN_A 2NOT_A 1AE7_A 3ELO_A 1GH4_A 1P2P_A 1GP7_A 1MH2_A 1MH7_A 1YXH_A 1LN8_A 1POA_A 1A3D_A 1A3F_A
33 33 36 36 40 41 52 53 54 55 57 57 67 78 84 89 92 95 100 100
1GOD_A 33
1SV3_A 33
1AOK_A 36
1BK9_A 36
1M8R_A 40
1VIP_A 41
1BUN_A 52
2NOT_A 53
1AE7_A 54
3ELO_A 55
1GH4_A 57
1P2P_A 57
1GP7_A 67
1MH2_A 78
1MH7_A 84
1YXH_A 89
1LN8_A 92
1POA_A 95
1A3D_A 100
1A3F_A 100
2. Cut/paste homologous domain and generate ddi 1bis and ddi 2bis
3. Superimposition of ddi 1bis and ddi 2bis.
4. check the clashing residues at the interface between dom1 and dom2.

Figure S3

Analysis of backbone clashes for residues at the interface. A. Distributions of the analyzed complexes. The upper limit of the box indicate the upper quartile of the distributions and the end of the whiskers shows the higher data point within one interquartile range of the upper quartile of the distributions. B-E. Histograms of values obtained for the percentage of interface residues with backbone clashes for Complex 2, 3, 4 and 5.

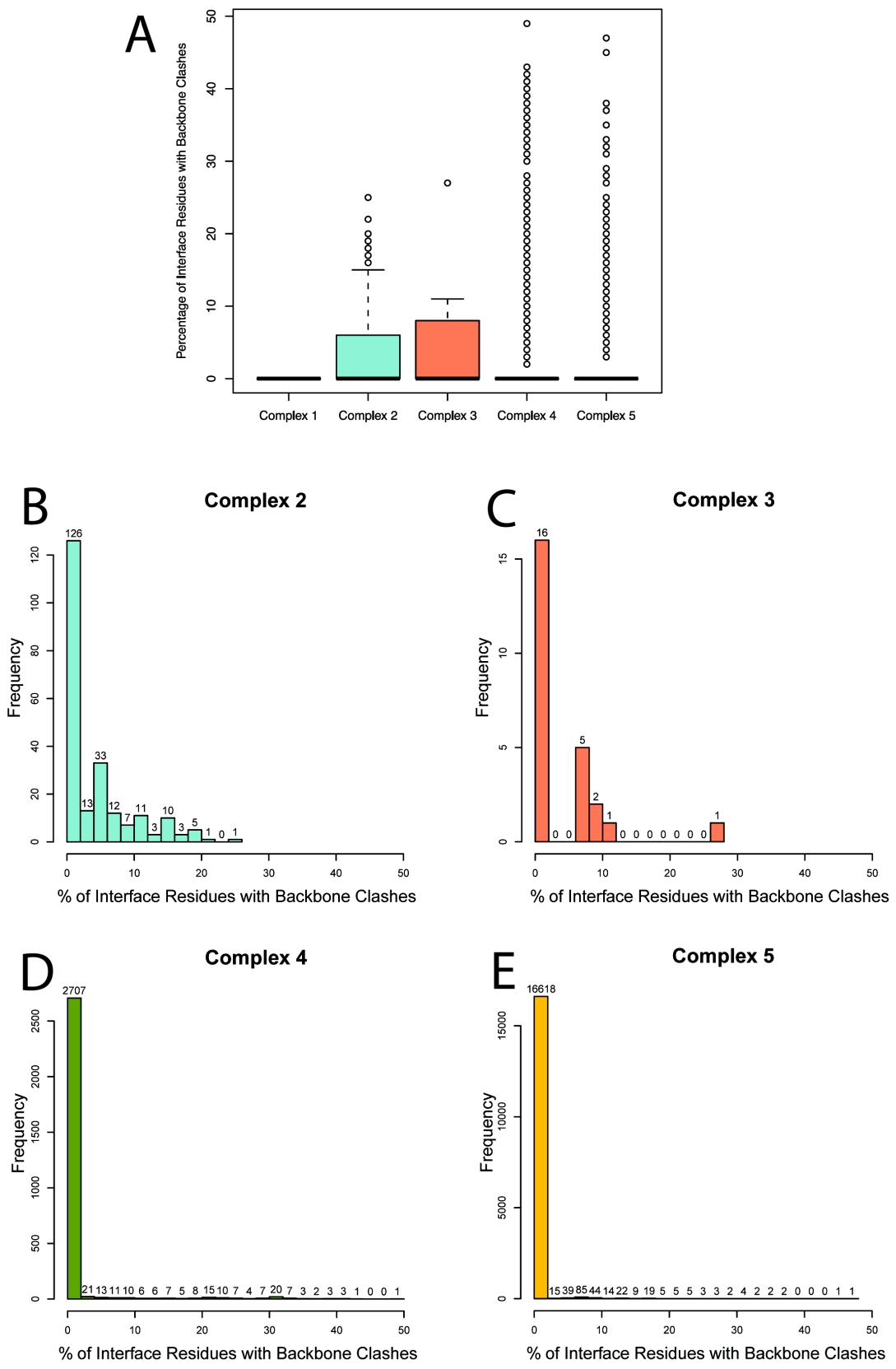
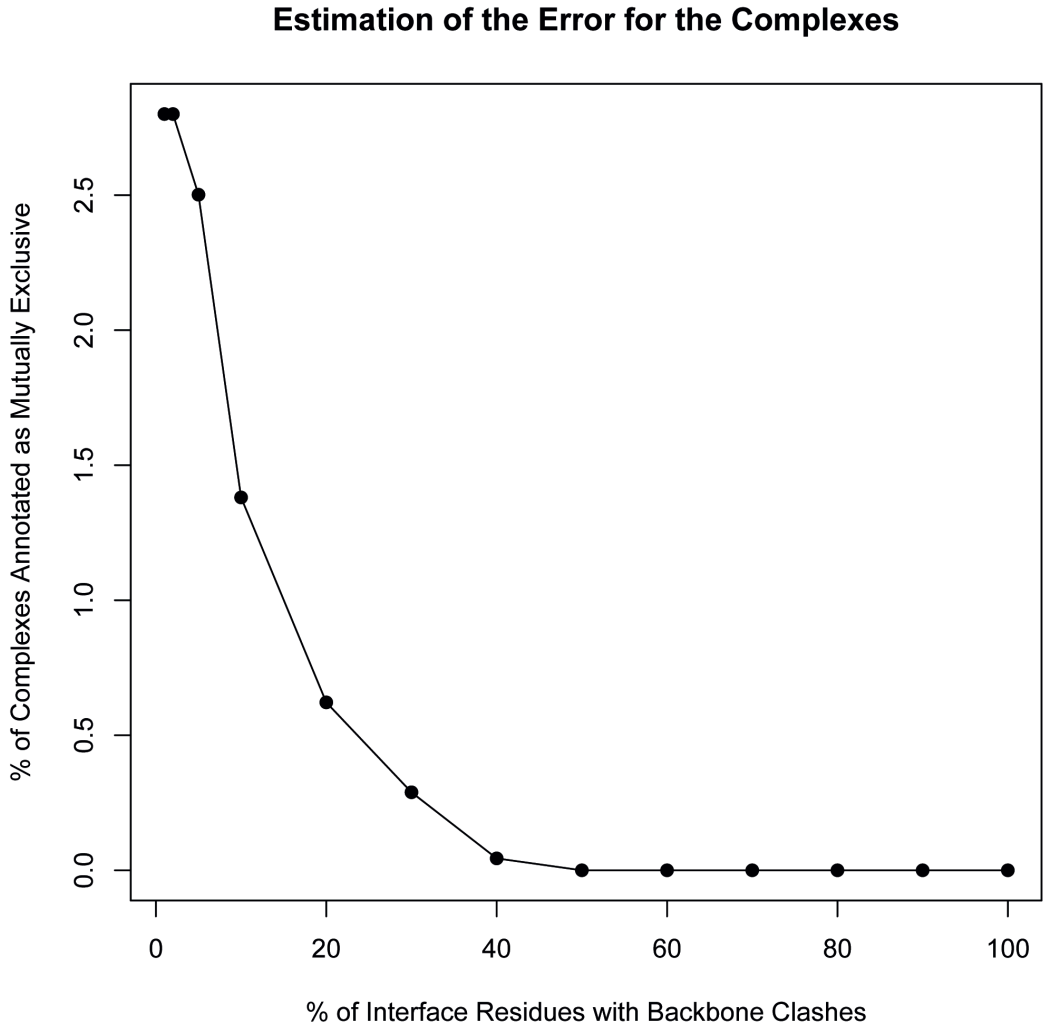

A
Percentage of Interface Residues with Backbone Clashes
0
10
20
30
40
50
Complex 1
Complex 2
Complex 3
Complex 4
Complex 5
B
Complex 2
Frequency
126
13
33
12
7
11
3
10
3
5
1
0
1
% of Interface Residues with Backbone Clashes
C
Complex 3
Frequency
16
0 0
5
2
1
0 0 0 0 0 0 0
1
% of Interface Residues with Backbone Clashes
D
Complex 4
Frequency
2707
21 13 11 10 6 6 7 5 8 15 10 7 4 7 20 7 3 2 3 3 1 0 0 1
% of Interface Residues with Backbone Clashes
E
Complex 5
Frequency
16618
15 39 85 44 14 22 9 19 5 5 5 3 3 2 4 2 2 2 0 0 0 1 1
% of Interface Residues with Backbone Clashes

Figure S4

Evaluation of the superimposition error. According to the different thresholds for interface clashing residues, we show the percentage of reconstructed complexes being filtered out. At the lowest threshold (1% of clashing residues), 2.8 % of all the cases are evaluated as being mutually exclusive.



3.3 Assessing structural interaction predictions

3.3.1 Introduction

There is an increasing gap between the amount of biochemical interaction data produced and the number of crystal structures of complexes that are being solved. In this project, we are aiming at filling this gap by extrapolating the detailed information provided by three-dimensional structures of proteins to large assemblies that constitute the protein interaction networks. In addition to describe the protein interactions at the atomic level, using structural data at the network level would add the important feature of competition and hence dynamics into the classical and static edges-and-nodes way of depicting protein interaction networks, by stating which interactions could happen at the same time and the ones that could not. We have developed SAPIN, a pipeline to analyze protein interaction networks from a structural point of view which consists of two main steps: first, pairwise protein interactions are treated to find a suitable structural template, and second the structurally predicted interactions are analyzed to identify the compatible and mutually exclusive interactions at the network level. The first step includes a direct search in the Protein Data Bank (PDB), to retrieve the structural data of a protein interaction when available or a structure of close homologous interacting proteins (i.e. sharing 70% of sequence identity). Additionally, based on the fact that globular proteins usually interact through domains and that similar domains usually interact in the same way (Aloy *et al.*, 2003), we integrated in our pipeline the search of potential structures of domain domain interactions (DDIs), provided by the 3DiD database (Stein *et al.*, 2011a). However, a given pair of interacting domains can exhibit multiple topologies of interaction (Kim *et al.*, 2006b) and the selection of the most appropriate template among the possible ones was consequently required. A first approach to tackle this issue has been to develop a scoring system taking into account the structural variability among domain families. As described in section 1.2.3, sequence-based domain classifications are usually derived from multiple sequence alignments and their members can exhibit low sequence identity. At the structural level, while the secondary structure elements arrangements are usually conserved in the domain folds, there can be more variation in loops or disordered

regions. This variability can thus affect a binding interface by making it prone or not available to form interactions.
Another strategy to select the most suitable template within a set of DDI structures was to use InterPreTS, which evaluates the pairwise contact between the interface residues base on empirical potentials (Aloy & Russell, 2002).
Structural data provided by the PDB may contain interactions that are actually artifacts generated at the crystallization step of the structure determination process (i.e. crystal packing) and are consequently not biologically relevant. As described in section 3.2, we have processed the structures of DDIs from 3DiD to filter out these non-biological contacts. Surprisingly, in some cases InterPreTS scores a crystallographic interface higher than a biological interface.
We have compiled positive (i.e. interacting proteins) and negative (i.e. non-interacting proteins) data sets to evaluate the performance of (1) our scoring system based on the structural variability of domains and (2) the version of InterPreTS which disregards the crystallographic artifacts.

3.3.2 Benchmarking structural templates as model for protein interactions

To benchmark our structural predictions of protein interactions, we compiled several reference sets with data from various resources: large-scale and low-scale experiments, structural predictions based on interface modeling and literature. They are summarized in table 3.1. For each type of data, there is a set of binding proteins and a set of non-binding protein, corresponding to a positive and negative control, respectively.

Yeast Two-Hybrid Data

Protein interactions can typically be identified by two types of experimental methods: Tandem Affinity Purification tag (TAP tag) and Yeast Two-Hybrid (Y2H) assays, capturing either stable protein complexes or binary physical interactions, respectively (these methods are discussed in section 1.2). The fact that Y2H assays capture binary interactions makes the protein interaction data coming from such experiments eligible to assess structural modeling. Another advantage of such data is the large amount of protein interactions

	Binding Set	Non Binding Set	Reference
Yeast Two-Hybrid	1500	1500 (X5)	Rual *et al.* (2005)
Ras-Effector Predicted Interactions	99	336	Kiel *et al.* (2007)
Ras-Effector Experimentally Tested Interactions	28	16	Kiel *et al.* (2007)
Literature	92	92	Braun *et al.* (2009)

Table 3.1: **Data sets used to evaluate the structural predictions** -

identified due the use of this technique a high-throughput fashion. We data provided by Rual *et al.* (2005). They tested pairwise interactions among the products of around 8100 human open reading frames (ORFs), and detected around 2800 interactions. Out of this data, we generated a first dataset, made of the interactions identified in this study (named "y2h_binding") containing 1500 interactions. Moreover, we extracted 'negative' sets from this data, by randomly extracting sets of 1500 interactions out of the ORFs that have been tested in the assay but for which no interaction has been detected. This is in fact another advantage of using Y2H data for benchmarking dataset, because it provides negative interactions. We generated 5 datasets (named "y2h_nonbinding") out of this negative data to ensure that the randomization had as a low effect as possible on the results. The advantages of using this data are 1) the identified interactions are binary, physically interacting with each other and 2)the negative set is made of proteins that have been tested in the system and have been found not to interact.

Ras-Effector interaction data

Kiel *et al.* (2007) used a structural approach which restricts the focus to the residues directly involved in the interaction. For a given interaction, all the elements (i.e. secondary structure elements and loops) that are not involved in the complex formation are excluded from the template. The side chains of the remaining residues in the template are then substituted *in silico* by the

side chains of the corresponding residues in the sequence of the proteins to be modeled. They predicted 20 Ras proteins in complex with 50 Ubiquitin-like domains, and the resulting network showed very high accuracy in distinguishing between binders (in 78% of the cases) and non-binders (in 83% of the cases) when compared to pull down experiments and binding affinities from literature. We used this binding and non binding data as a dataset, by distinguishing on one hand the predicted interactions (named "ras_pred_binding" and "ras_pred_nonbinding", containing 99 and 336 interactions, respectively) and on the other hand the interactions that have been experimentally tested (named "ras_exp_binding" and "ras_exp_nonbinding", containing 28 and 16 interactions, respectively).

Literature-based data

Braun *et al.* (2009) compiled a positive reference set (PRS) out of interactions for which there was more than one peer-reviewed publication in manually curated databases. They also generate a random reference set (RRS) considered as a negative set, since two proteins picked randomly among the proteome are less likely to interact with each other. These sets ("PRS_vidal" and "RRS_vidal") both contain 92 interactions. Interaction data contained in databases can be considered as valuable especially in the cases where there are several pieces of evidence, and when the experimental method used is a "low-scale" approach.

The advantage of using several sources for benchmarking sets is that it allows to compensate for biases inherent to a given type of data. For example, Y2H data is often seen as noisy because it is estimated to contain false positives and false negatives, but it is convenient for our purpose because it contains binary interactions and provides a large amount of data. Given the low coverage when dealing with structural interactions, a reasonably sized dataset is thus of great statistical value. Conversely, the literature-based datasets are rather small, and even if it is considered as being of higher quality, it could result from the mining of the research article potentially describing a crystal structure. The Ras-Effector data sets are similarly

small for the predicted interactions and even smaller for the experimentally tested interactions, but the latter are considered of high quality, because they result from experimentally validated predictions.

3.3.3 Selection of an interaction template based on the structural variability of domains

Domain database classifications and structural similarity

The diversity of classification methods of protein sequences into domain families has led to a large variability of the description of a given sequence in a given database. For example, some methodologies perform a simple sequence similarity search, while others are based on HMM profiles generated from multiple sequence alignments and some take into account functional features. At the structural level, a domain is characterized by the fact that it can form a stable fold independently from the rest of the protein. Overall, the sequence-based domain classifications agree with the fold definition, but their domain signatures can differ. A striking example is the Ras domain family. To study the structural variation of the Ras domain, we performed a pairwise structural alignment of its members as classified by PFAM (Finn *et al.*, 2010b) (Figure 3.1). We identify structurally similar subfamilies which correspond to four families described in the sequence-based SMART database (Schultz *et al.*, 1998b), i.e. Rho, Ras, Ran and Rab.
InterPro is a meta-database that integrates protein domain families from diverse databases sources (Hunter *et al.*, 2009). In this integration effort, InterPro also indicates the family/subfamily relationships between the various database entries. For example, regarding the previously described example of the Ras family, InterPro classifies the SMART Rho, Ras, Rab and Ran families as subfamilies of the PFAM Ras domain family. The tree covering the relationships derived from all the integrated databases in InterPro for the Ras domain family is shown in figure 3.2 A.

A score based on the functional annotation of protein domains

The different domain classification methods can reflect a structural variability among the members of a domain family. This structural variation between domain families

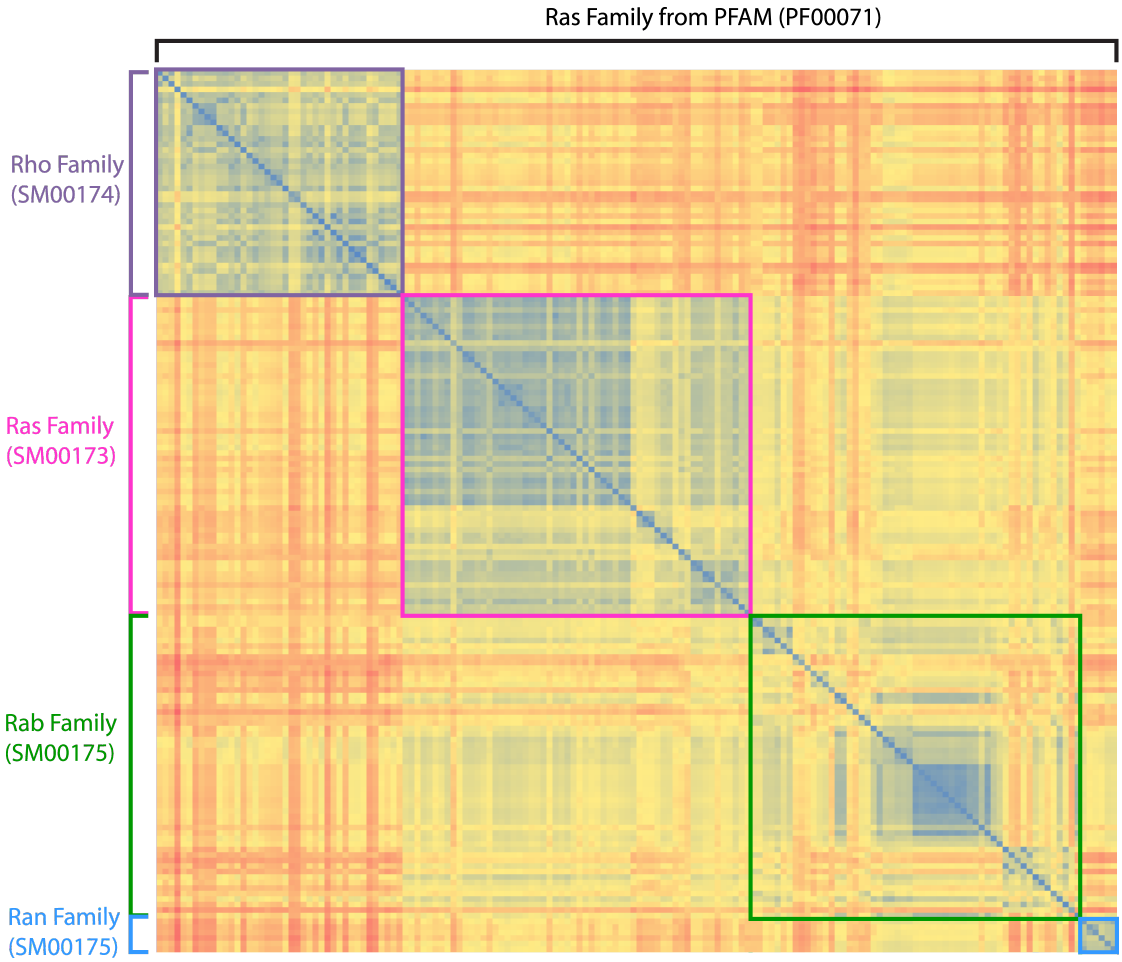


Figure 3.1: **Pairwise structural alignment of the Ras family members** - The Ras family members from PFAM have been structurally superimposed using Dali (Holm & Park, 2000). A high score (blue) indicates that the pairs are structurally similar, while a low score (red) is obtained the tested pairs that present a structural variability.

implies a different conformation and can consequently have an effect on the way an interaction between two domains is mediated. This principle served as a basis for implementing a scoring system, which objective is to select the most appropriate template out of a pool of interacting domain pairs. The several steps for the implementation of the scoring system are the following:

1. The domain interaction structures from 3DID are mapped on the InterPro classification: each domain is assigned with the lowest database entry in the InterPro classification tree previously described. Similarly, the protein

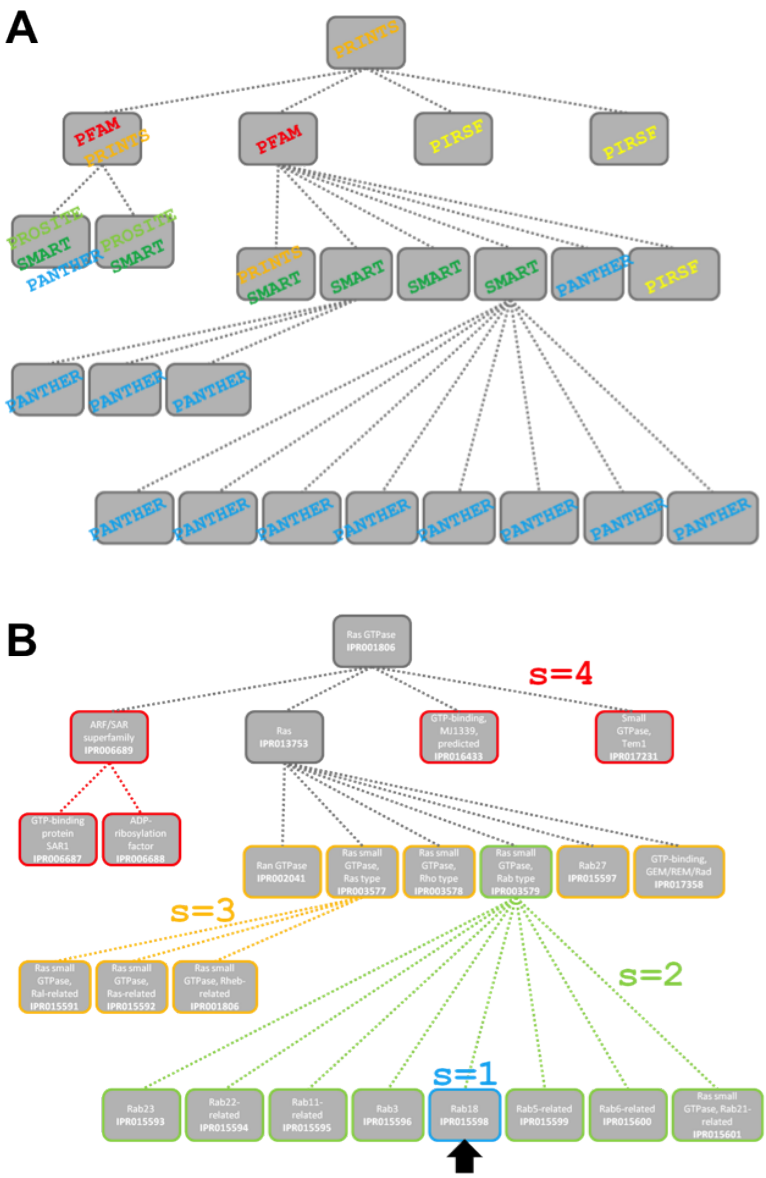


Figure 3.2: **Relationships between the domain classification methods as defined by InterPro** - (A) Representation of families/subfamilies relationships for the Ras family, according to InterPro. (B) Illustration of the InterPro tree-based scoring method. Given two domains, if the first domain corresponds to a given InterPro entry (black arrow), the algorithm searches how far from this entry the second domain is in the classifiaction tree: in the same group (blue), in another group of the same level (green), in another group of the same family (yellow) or in a higher branch of the tree (red). A score is assigned accordingly to 1, 2, 3 or 4 respectively.

sequences from the interaction to be modeled are assigned with InterPro entries.

2. We group the parts of each family tree to implement the scoring system: entries belong to the same group if all of them are dead ends and none of them have a child. If, at a given level in the tree, one of the entries has the child, then all the entries of this level with no child belong to a different group. The algorithm then searches the tree to identify how far a domain assigned to the query protein is from the domain assigned to the structural template: in the same group, in another group of the same level, in another group of the same family or in a higher branch of the tree. The score is assigned accordingly to 1, 2, 3 or 4 respectively (Figure 3.2 B).
3. Finally, we compute the score evaluating the domain interaction template. The resulting score takes into account the number of families encompassed in a given InterPro tree. It ranges from 0 to 1 and the higher the score, the better the template, i.e. the domains from the structural template are close to the domains from the interaction query.

Evaluation of the scores

We used the positive and negative reference sets generated from Y2H data (see section 3.3.2) to evaluate the biological relevance of this score. As not all the InterPro entries have family/subfamily relationships, we performed the evaluation by (1) taking into account all the domain families regardless of their organization in trees in InterPro and (2) leaving out the domain families for which no relationships have been described (i.e. containing only one InterPro entry). Figure 3.3 shows the distributions of the scores obtained for the positive set and the negative set (namely "y2h_binding" and "y2h_nonbinding", respectively) in these two tests. In both cases, the distributions for the negative and positive sets are very similar, and centered at the score of 0.7. These results suggest that this approach does not allow an accurate evaluation of domain interaction structures, and that the information provided by the various domain classification methods and their relationships remains insufficient to distinguish correctly between binding and non-binding proteins.

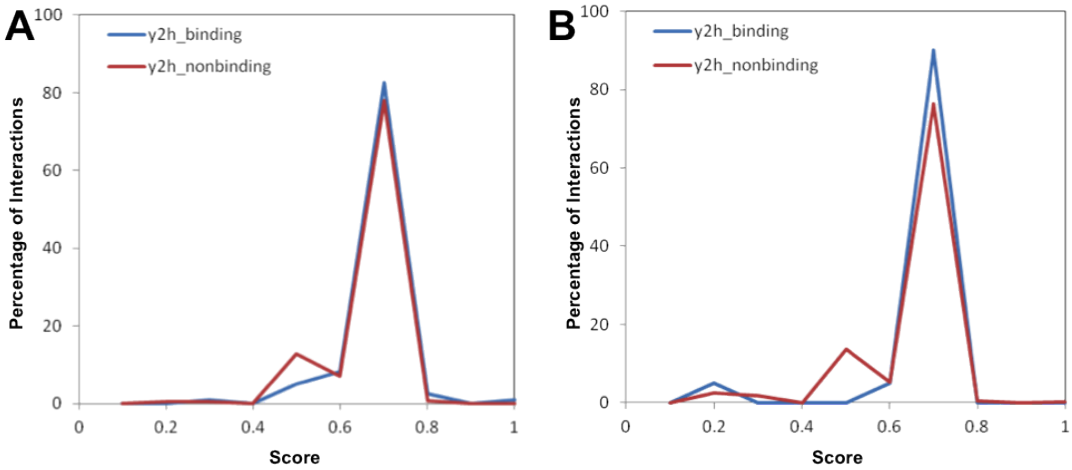


Figure 3.3: **Evaluation of the score based on the InterPro classification** - The positive (in red) and negative (in blue) reference set coming from Y2H data are scored according to the InterPro tree-based method, taking into account all the InterPro families (A) and excluding the families for which there is no relationships (B).

Despite the structural variability observed among the diverse domain families, this information alone can not help selecting a favored way of interacting.

3.3.4 Selection of a structural template using InterPreTS

Another strategy to evaluate and select a structural template is to take into consideration the interaction interface. We used InterPreTS (Aloy & Russell, 2003), a method that uses empirical potentials, to describe how well two sequences fit into a structural template of an interaction between homologous proteins. This method has been applied successfully to predict the specificities of large domain families (Aloy & Russell, 2002). An InterPreTS Z-score above or equals to 2.3 indicates a significance of the prediction of 99%. However, InterPreTS sometimes gives a low score to a biological interface and could score highly an interface known to be a crystallographic artifact. Such an example is shown in figure 3.4, where InterPreTS scores better a crystallographic artifact than the two biological units. In our framework SAPIN (described in section 3.2), we cope with these potential cases by

filtering out the non biologically relevant interfaces.

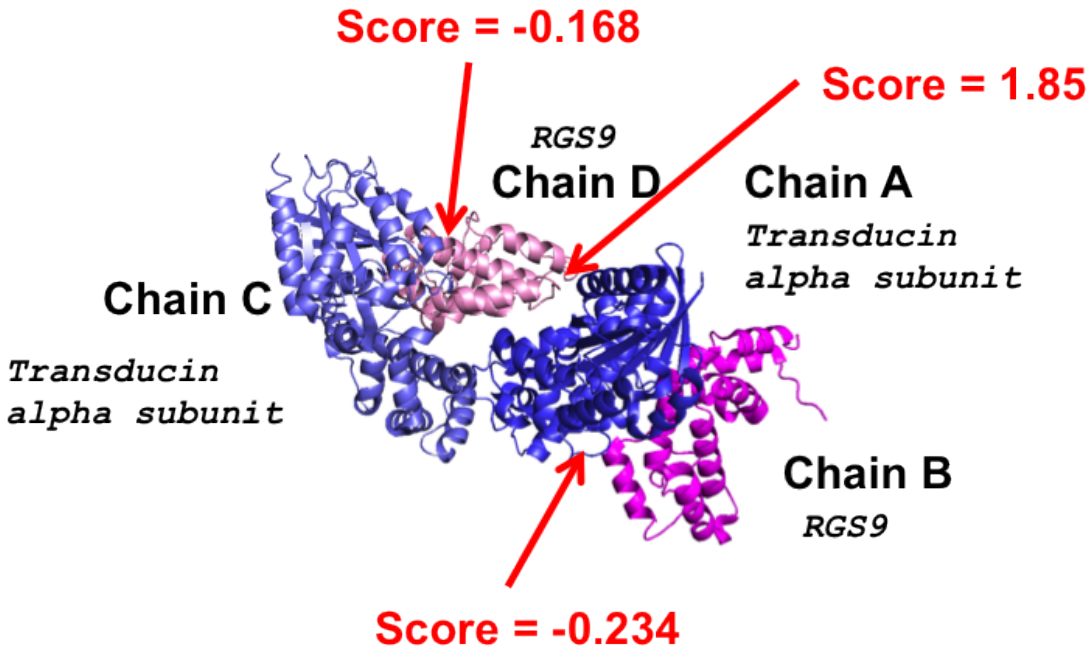


Figure 3.4: **Example of InterPreTS scoring** - The PDB entry 1fqk contains two copies of the alpha subunit of the G Protein transducin in complex with RGS9.

Effect of randomization and crystallographic artifacts on InterPreTS scores

We further evaluated the prediction performance of InterPreTS, using the positive and negative reference sets from Y2H data, under different conditions.

1. As InterPreTS scores are based on randomly generated sequences, the resulting score for a given pair of sequences can vary from one run to another. We thus ran InterPreTS ten times for each interaction and computed the average and standard deviation. When compared to a one-run InterPreTS, we could observe a slight variation in the distribution of the scores for both positive and negative sets, but this effect does not seem to affect the separation between the two distributions (Figure 3.5 A and B), and thus the distinction between binding and non-binding data.

2. We further evaluated the influence of disregarding the results of predictions using interfaces have been annotated as being crystallographic artifacts. The resulting distributions exhibit a clearer separation between binding and non-binding data, especially the score distribution for the negative set being shifted towards lower values, whereas the distribution for the positive set remains at the same level than the distribution obtained previously (Figure 3.5 C).
3. We performed an extra test taking into consideration the interactions for which we could identify at least one domain interaction in 3did. The resulting distributions are equivalent to the previous ones, in terms of separation. This corresponds to our approach of combining the search in 3DiD and the scoring by InterPreTS (Figure 3.5 D). The distributions in this case are more homogeneous for both sets and suggest a slight improvement in distinguishing between binding and non-binding proteins.

Evaluation of InterPreTS predictions

To extensively evaluate the predictions provided by InterPreTS, we used the reference sets described in 3.3.2 and summarized in table 3.1.

The structural coverage varies for each type data in the reference sets (Figure 3.6). The interaction sets coming from Y2H data are the ones with the lowest coverage (10% for the positive set and around 1% for the negative sets, figure 3.6 A). The structural coverage for the positive set is in agreement with the estimated number of structural interactions for the human interactome (Stein *et al.*, 2011b). InterPreTS predicts an interaction for more than 50% of the interactions from the Ras-Effector interaction positive data sets (predicted and experimentally tested) and for 45% for the negatively predicted Ras-effector interactions (Figure 3.6 B). Surprisingly, up to 80% of experimental non-binding Ras-Effector interactions are found to have a structural template (Figure 3.6 C). As for the literature-derived interaction set, the coverage reaches 30% for the positive set and no structural template has been found for the negative set (Figure 3.6 D).

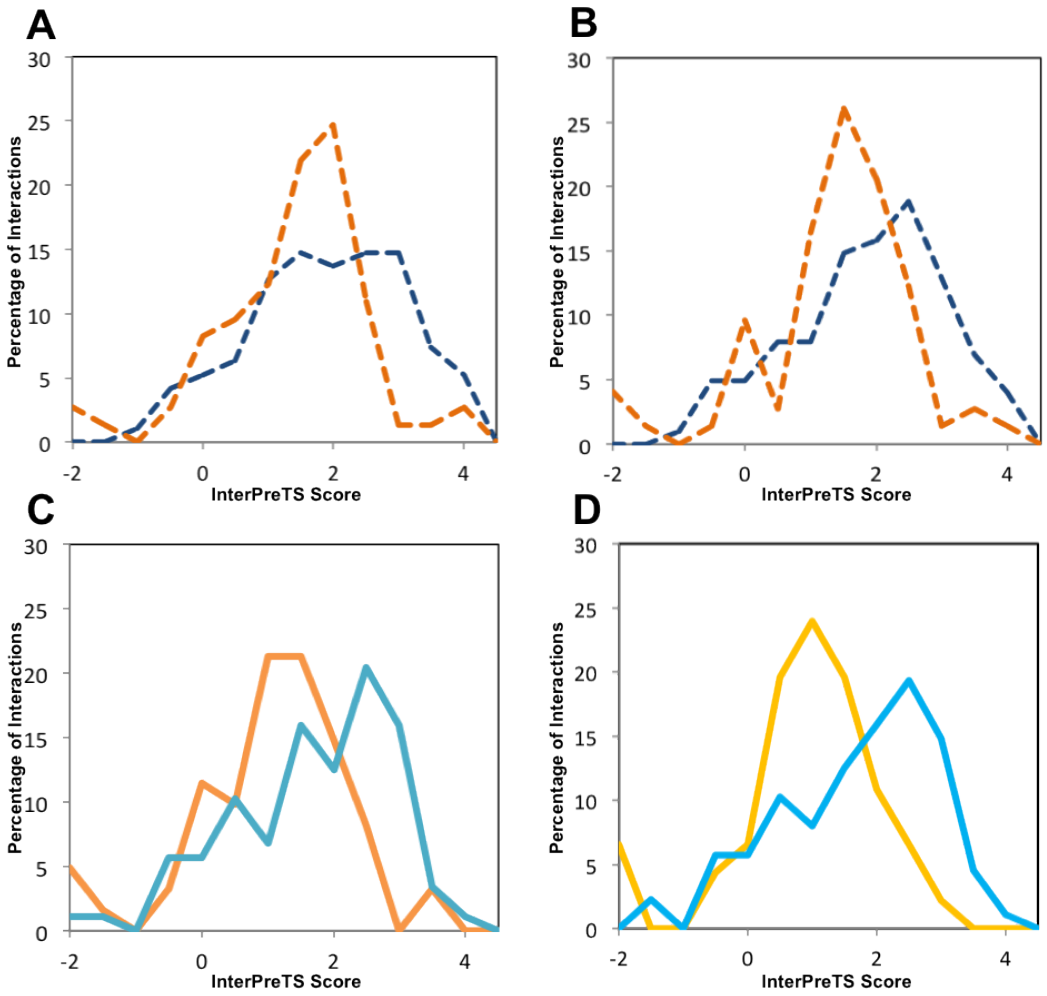


Figure 3.5: **Distributions of InterPreTS scores for Yeast Two-Hybrid reference sets** - InterPreTS predictions have been evaluated under diverse conditions for the positive reference set (orange) and negative reference set (blue). (A) InterPreTS has been ran once on each interaction. (B) The predictions have been repeated ten times for each interaction and the the average scores are computed. (C) The crystallographic artifacts have been filtered out. (D) InterPreTS has be used only on interactions for which there is an available structure of domain interaction.

The distributions of InterPreTS Z-scores for the eight reference sets tested are shown in figure 3.7 (top). Except for the literature derived sets (because no structural template has been found for the negative set), all the distributions exhibit a similar

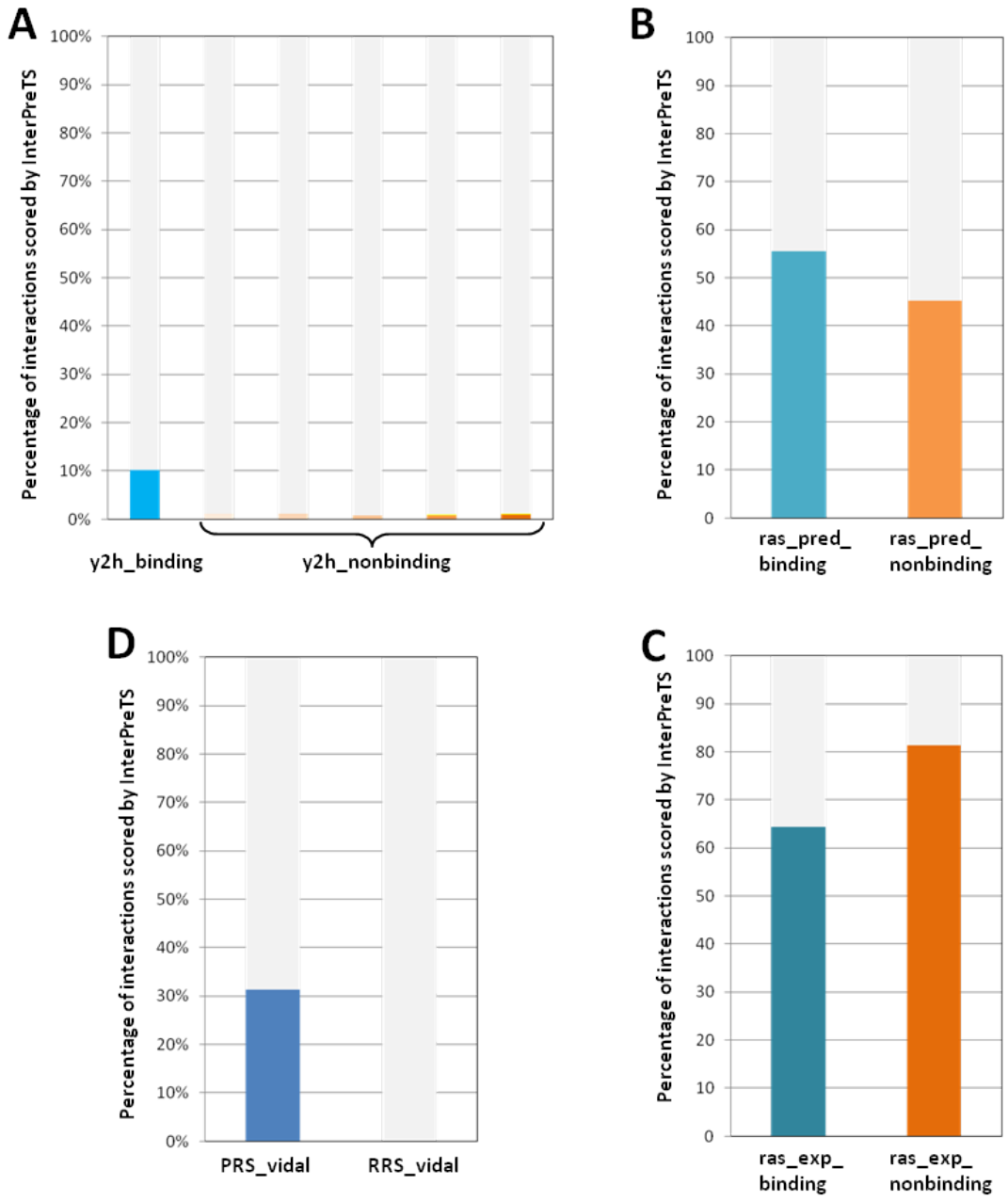


Figure 3.6: **Structural coverage of the different reference sets used** - The structural coverage for the positive and negative reference sets are shown in blue and orange respectively for the four types of data: (A) Yeast Two-Hybrid (B) Ras-effector predicted interactions (C) Ras-effector experimentally tested interactions (D) Literature-derived interactions.

trend: the distributions of scores for the positive and negative sets are clearly shifted, which suggests that the scores allow the distinction between binding and non-binding data. Interestingly, the variation observed in terms of structural coverage (i.e. a low coverage for the Y2H data and a high coverage for the Ras-Effector interaction data) does not affect the score distributions.

The sequence identity between the structural templates and protein queries seems to have a relatively limited effect on the score distributions (3.7, bottom). In fact, most of the proteins from the various negative sets share less than 50% identity with their structural template. In contrast, many proteins from the positive set share higher sequence identity with their templates, which is often associated with a high InterPreTS score. This only applies to a limited number of cases, as the majority of proteins sharing between 30 and 50% have a relatively high score.

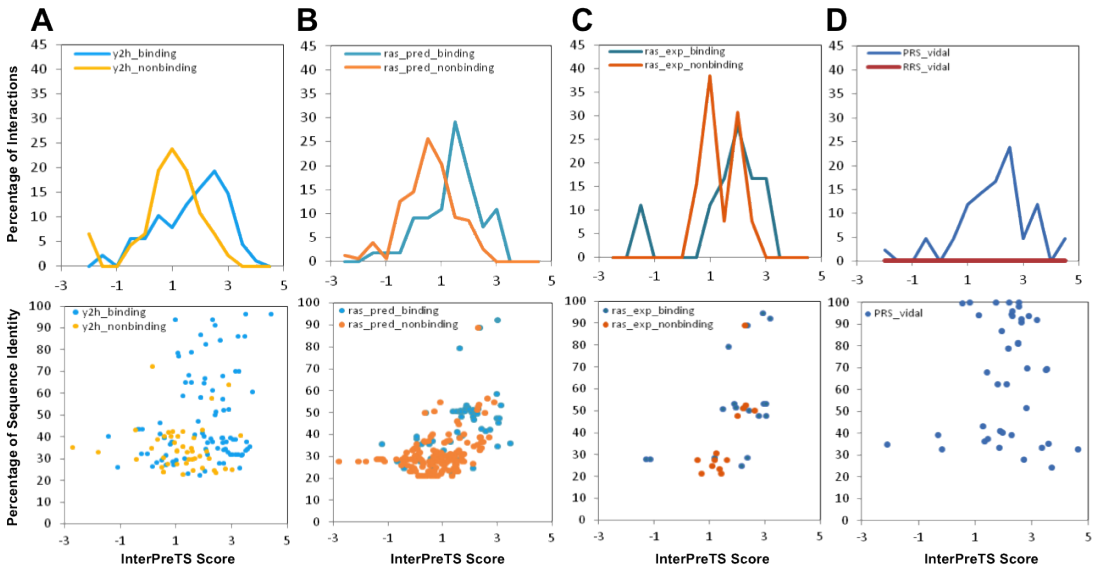


Figure 3.7: **Distribution of InterPreTS scores for each type of reference set** - InterPreTS predictions have been evaluated using positive references sets (orange) and negative reference set (blue). The distirbutions of InterPreTS scores are shown at the top and the sequence identity between query and template sequences are shown at the bottom for the four types of data: (A) Yeast Two-Hybrid (B) Ras-effector predicted interactions (C) Ras-effector experimentally tested interactions (D) Literature-derived interactions.

The receiver operating characteristics (ROC) curves of the reference sets illustrate the tradeoff between true and false positive rates. The ROC curves for results

from the predictions for the Y2H and Ras-Effector interaction data are shown in Figure 3.8. The prediction performance is higher for the Ras-Effector predicted interactions. This suggests that both prediction methods seem to be in agreement. The performance on the Y2H data is similar to the one observed for the Ras-Effector experimentally tested interactions.

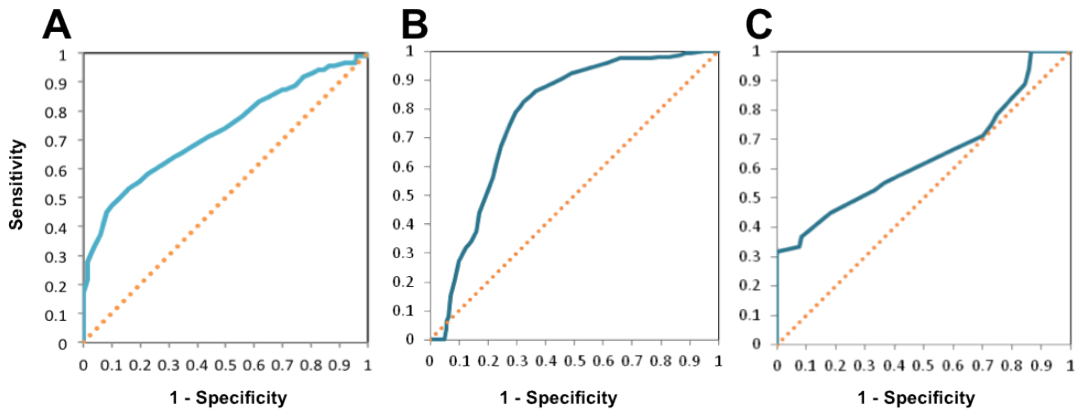


Figure 3.8: **Evaluation of the prediction performance** - The receiver operating characteristics (ROC) curve shows the tradeoff between true and false positive rate for the Y2H (A), Ras-Effector predicted (B) and exprimentally tested (C) interaction reference sets.

InterPreTS has been successfully used to predict protein interactions (Aloy & Russell, 2002). However, it can give high scores to interfaces that are actually crystallographic artifacts. Filtering out such non biologically relevant interactions from 3DiD has shown that around 27% of the domain interaction interfaces were considered as being non functional interactions. Using our reference sets, we evaluated InterPreTS predictions we could (1) improve the structural predictions made by InterPreTS and (2) validate the use of InterPreTS to select the most suitable template out of a pool of structures of interacting domains.

CHAPTER

FOUR

DISCUSSION

The objective of this dissertation is to focus on one of the most challenging aspects in systems biology: truly understand protein interaction networks on a functional level. Classical networks represent a rather static picture of biological processes and therefore we proposed a more dynamical view by introducing the missing feature of compatible and mutually exclusive interactions. We argued that knowledge of interacting interfaces provided by structural information can not only help determining the possible binary interactions taking place within large complexes but can also contribute to the elucidation of competition between binding proteins, by discriminating compatible and mutually exclusive interactions. We focused our effort on combining domain-domain and domain-peptide interaction predictions with experimental information, and proposed an automated method to facilitate such a structure-based dynamic view of protein interaction networks.

In section 3.1, we combined structural information with literature mining and proteomics data to construct the protein interaction network associated with rhodopsin. The further annotation of proteins according to their physiological functions permitted the decomposition of the network into functional sub-modules. We applied a structural analysis to predict compatible and mutually exclusive interactions, which allowed the annotation of the connections among and between sub-modules using logical gates "AND" and "XOR", respectively (Campagna *et al.*, 2008). This functional and structural analyses suggested two novel links from rhodopsin towards (1)

cytoskeleton dynamics regulation and (2) vesicle trafficking in order, which were partially supported by experimental data.
Taken together, these results showed structural information combined with experiemtal data is useful to derive new functional insights. Despite a low structural coverage in the Rhodopsin protein interaction network, we were able to gain insight about potential new routes activated by rhodopsin.
Using the structures of domain interactions as template to model protein interactions, we were able to increase the structural coverage of the rhodopsin-associated network by 50%. These results encouraged the automatization of this structural approach to facilitate its application to other systems, where experimental information is available.

In section 3.2 of this thesis, we have developed a pipeline, SAPIN, for the structural analysis of protein interaction networks in an automated way. The objectives of this framework were: (1) to structurally predict binary protein interactions within protein complexes and (2) to identify the possible mutually exclusive and compatible interactions.

Our approach assumes a rigid backbone, without any conformational change between the unbound and bound forms of the proteins involved. This corresponds to the lock-and-key principle first observed in enzymatic reactions by Fischer in 1894 (Fischer, 1894). Stein *et al.* (2011d) have recently measured the flexibility of protein domains upon association and showed that in 65% of the cases the interaction indeed did not lead to any conformational change. In 13% of the cases the interactors followed a conformation selection model (Goh *et al.*, 2004), where their bound conformation is encountered at regular intervals during exploration of the unbound state. Therefore, structural protein interaction methods could benefit from using flexible backbone methods, which is computationally more exepensive but becomes nevertheless more feasible with the constant increase in computational power. Recently, another protocol was proposed to structurally predict protein interactions at a large-scale level (Tuncbag *et al.*, 2011). This method is based of the fact that there is a limited number of naturally occurring architectural motifs. The

authors use PRISM (Ogmen *et al.*, 2005), a resource containing these structural motifs, which is searched by structural superimposition in order to identify the hot spots of the target proteins and then for potential complementary between them. This approach is complementary with ours since it performs a flexible docking for monomeric structures of proteins. Since each method covers a distinct interaction type model (i.e. the lock-and-key model and the conformational selection model), we can speculate that combining these approaches would lead to an increase of the structural coverage.

We have focused our structural approach on domain interactions rather than protein interactions with the objective of increasing the structural coverage. However, the domain interactions predicted as being compatible might not reflect the reality in the case of multi-domain proteins. In fact, the spatial arrangement of one or several domains from the same peptide chain can impede the simultaneous binding of a third protein. Structures of full-length proteins if available could thus help to predict more accurately the compatible binding of domain-based interaction models. Integrating the structural neighborhood of a domain interaction, for example by superimposing the structure of the full-length protein, could thus improve the accuracy of compatibility predictions. This would in principle be limited to the cases in which proteins have been structurally resolved, but the fact that the structural coverage is higher for proteins than for protein interactions (Stein *et al.*, 2011c) is encouraging even though this would increase the complexity of computational approaches and their execution time.

Moreover, our domain-based approach does not take into account the fact that domains can be repeated within a protein. This could lead to a misinterpretation of interactions predicted as mutually exclusive. Indeed, if two domains A and B are predicted to bind to the same surface of a third domain C and if the latter is repeated (i.e. C1 and C2), then A and B might actually be able to bind simultaneously to respectively C1 and C2 and vice versa. This complex problem is difficult to address within the scope of our approach, and would require a careful manual inspection combined with extra contextual information about the interactions.

Once the interactions have been established to be mutually exclusive, a logical step would be to push the analysis towards the concept of conditional interactions. Post-translational modifications can act as switches, allowing or impeding the bind-

ing of proteins. Thus, if two proteins are predicted to bind a common surface, the knowledge of such switches can help establish the order in which they should bind. This could add a time dimensionality into networks, but in turn would increase the complexity.

It is obvious that competing interactions can be identified based on the knowledge of interacting interfaces. However, this information needs to be completed with quantitative data about the concentration of proteins competing for the same interface. Given two proteins A and B competing for binding with a third protein C, protein A is present in a concentration 100-fold higher than protein B, the interaction between A and C will be preferred. Similarly, the localization is important to take into account. For example, in the human interactome, some proteins being expressed in tissue-specific manner may never have to compete with proteins expressed in different tissues.

In section 3.3, we described the strategies we used to cope with the two main issues related to the use of 3D structures of interacting domains: (1) the atom coordinates, as provided by the PDB, often contains interface that are non biologically relevant as they result from the crystallization process and (2) a pair of interacting domains can exhibit multiple topologies of interactions.

First, we annotated crystallographic artifacts and biological interfaces among the interfaces within the structural data coming from 3DiD. We found that a non negligible portion (around 27%) of the interactions were actually considered as non biologically relevant by our filtering. Further, the evaluation of InterPreTS scoring showed that the predictions were improved after the interfaces considered as crystal packing had been filtered out.

Second, we implemented a scoring system aiming at selecting the best template out of a pool of domain interaction structures exhibiting different interaction topologies. This was based on the observation that domains could have a relatively important structural variability within a domain family and that domain classifications can provide different domain definitions. We used the family/subfamily relationships between these domain signatures to try and explain the diverse topologies ob-

served by the structural variability of domains. However, this approach did not allow relevant structural prediction, as it gave similar scores to binding and non-binding data, which could be partly explained by the fact that a vast majority of domain families are not interconnected, thus the relational approach was limited to a small number of cases.
An crucial step when developing a prediction method remains the validation. To test our methodology, we have used four data sources to generate positive and negative reference sets: Yeast Two-Hybrid data, predicted and experimentally tested Ras-effector interactions and data from literature. With these various datasets, we could cover many properties of interaction data: binary physical interactions, predicted and experimentally identified interactions, large- and low-scale experiments, high and low structural coverage. The evaluations carried out on the different sets showed similar trends in terms of predictions, with a slightly better performance on the predicted Ras-Effector interactions suggesting that InterPreTS predictions were in agreement with the interface modeling approach used to generate this data. However, the diverse reference sets exhibited different structural coverages. In particular, Ras-Effector interaction data had a surprisingly high coverage for the negative sets (up to 80% for the predicted interactions), while Y2H negative data coverage was in contrast extremely low. Yet, this had no effect on the distinction of binding and non-binding data. Taken together, these results suggest that the methods we used are robust to distinguish between binding and non-binding proteins, independently from the type of data.

Despite the initiative of structural genomics projects to produce 3D structures more effectively, the global coverage at the interactome level remains very low (around 10% for the human interactome as estimated by Stein *et al.* (2011b)) in comparison to the large number of sequence and interaction data. This lack of valuable data provided by structures represents the main limiting factor when dealing with protein structure modeling. Yet, the various computational methods available will be gradually improved as more structures become available. However, we have shown that even at a low structural coverage, the combination of experimental data

with structural information can help to gain functional and biological insight within a protein interaction network.

CHAPTER

FIVE

CONCLUSIONS

1. We built a comprehensive protein interaction network associated with the GPCR rhodopsin by combining structural information, proteomics data and literature mining and identified functional sub-modules. The approach used for the network reconstruction could be applied to any other cellular pathway. Further it opens the possibility of further experimental testing by scientists working on vision and vision-related diseases.
2. The protein interaction network we constructed served as a basis for the analysis of compatible and mutually exclusive interactions, based on previously structurally predicted interactions. This functional and structural interaction network suggested two novel routes, linking rhodopsin to (1) cytoskeleton dynamics and (2) to vesicle trafficking regulation, which were validated experimentally. This could serve a basis to explore further hypotheses of light-regulated dynamics involved in the protein transport in the ciliary region of Rods.
3. By mapping the mutations known to be involved in diseases on the structural models, we identified the core vision pathway as being more susceptible to be affected by diseases, suggesting that the high-end functional properties of the visual pathway may have been preferred over robustness throughout evolution, enabling single photon detection and multi-color vision.
4. We developed SAPIN, a tool to analyze protein interaction networks using structural information in an automatized way. SAPIN combines a classical

approach of comparative modeling with the search of more remote homologous structures as templates by using 3D structures of domain-domain and domain-peptide interactions. In addition, we integrated prediction methods of binding motifs and phosphorylation sites. SAPIN contains an algorithm which, given a set of structurally characterized or predicted interactions, evaluates if they are compatible or mutually exclusive. This method highlights the principle of competition, which is known to be important in signal transduction pathways.

5. We tested the structural prediction results provided by our framework, by using diverse positive and negative reference sets and evaluated the compatibility analysis by testing our algorithm on a large number of cases. These results enabled us to validate our methodology, since the structural prediction could distinguish between interacting and non-interacting data and could confidently predict the the mutually exclusive interactions.
6. SAPIN will be shortly accessible through a web server, where one will be able to submit a set of interactions and its related sequences. The results from the structural analysis are displayed in an interactive visualization tool, to facilitate the browsing of the predicted interactions and their compatibilities. Thus, our framework could be used to automatically analyze any protein interaction network.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

Aloy, P. & Russell, R.B. (2002). Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5896–901.

Aloy, P. & Russell, R.B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* (*Oxford, England*), **19**, 161–2.

Aloy, P. & Russell, R.B. (2004). Ten thousand interactions for the molecular biologist. *Nature Biotechnology*, **22**, 1193.

Aloy, P., Ceulemans, H., Stark, A. & Russell, R.B. (2003). The Relationship Between Sequence and Interaction Divergence in Proteins. *Journal of Molecular Biology*, **332**, 989–998.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J.P., Chothia, C. & Murzin, A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic acids research*, **32**, D226–9.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S.N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K. & Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic acids research*, **38**, D525–31.

Bader, G.D. & Hogue, C.W.V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nature biotechnology*, **20**, 991–7.

Bader, G.D., Betel, D. & Hogue, C.W.V. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, **31**, 248–250.

Barabási, a. (1999). Emergence of Scaling in Random Networks. *Science*, **286**, 509–512.

Beltrao, P. & Serrano, L. (2005). Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. *PLoS computational biology*, **1**, e26.

Bernauer, J., Bahadur, R.P., Rodier, F., Janin, J. & Poupon, A. (2008). DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics* (*Oxford, England*), **24**, 652–8.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., Smet, A.s.D., 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. *Nature Methods*, **6**, 91–97.

Brown, K.R. & Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics* (*Oxford, England*), **21**, 2076–82.

Bustos, D.M. & Iglesias, A.a. (2006). Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins. *Proteins*, **63**, 35–42.

Campagna, A., Serrano, L. & Kiel, C. (2008). Shaping dots and lines: adding modularity into protein interaction networks using structural information. *FEBS letters*, **582**, 1231–6.

Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. (2008). Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS letters*, **582**, 1171–7.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L. & Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic acids research*, **38**, D532–9.

Chatr-Aryamontri, A., Ceol, A., Licata, L. & Cesareni, G. (2008). Protein interactions: integration leads to belief. *Trends in biochemical sciences*, **33**, 241–2; author reply 242–3.

Chaurasia, G., Iqbal, Y., Hänig, C., Herzel, H., Wanker, E.E. & Futschik, M.E. (2007). UniHI: an entry gate to the human protein interactome. *Nucleic acids research*, **35**, D590–4.

Davis, F.P. & Sali, A. (2005). PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* (*Oxford, England*), **21**, 1901–7.

Deane, C.M. (2002). Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations. *Molecular & Cellular Proteomics*, **1**, 349–356.

Deeds, E.J., Ashenberg, O. & Shakhnovich, E.I. (2006). A simple physical model for scaling in protein-protein interaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 311–6.

Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J. & Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic acids research*, **39**, D261–7.

Enright, a.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.a. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y.V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.P., Duewel, H.S., Stewart, I.I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.L., Moran, M.F., Morin, G.B., Topaloglou, T. & Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular systems biology*, **3**, 89.

Fields, S. & Song, O.k. (1989). A novel genetic system to detect protein-protein interactions. *Nature*.

Finn, R.D., Marshall, M. & Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* (*Oxford, England*), **21**, 410–2.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R. & Bateman, A. (2010a). The Pfam protein families database. *Nucleic acids research*, **38**, D211–22.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R. & Bateman, A. (2010b). The Pfam protein families database. *Nucleic acids research*, **38**, D211–22.

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, **27**, 2985–2993.

Gandhi, T.K.B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J.D., Parmigiani, G., Schultz, J., Bader, J.S. & Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast , worm and fly interaction datasets. *Nature Genetics*, **38**, 285–293.

Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Höfert, 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.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi,

T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. & Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–6.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, a., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, a., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, a., Zhong, J., Stanyon, C.a., Finley, R.L., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.a., McKenna, M.P., Chant, J. & Rothberg, J.M. (2003). A protein interaction map of Drosophila melanogaster. *Science* (*New York, N.Y.*), **302**, 1727–36.

Goh, C.S., Bogan, a.a., Joachimiak, M., Walther, D. & Cohen, F.E. (2000). Co-evolution of proteins with their interaction partners. *Journal of molecular biology*, **299**, 283–93.

Goh, C.S., Milburn, D. & Gerstein, M. (2004). Conformational changes associated with protein-protein interactions. *Current opinion in structural biology*, **14**, 104–9.

Goll, J., Rajagopala, S.V., Shiau, S.C., Wu, H., Lamb, B.T. & Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics* (*Oxford, England*), **24**, 1743–4.

Gould, C.M., Diella, F., Via, A., Puntervoll, P.l., Gemünd, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J.C., Chica, C., Seiler, M., Davey, N.E., Haslam, N., Weatheritt, R.J., Budd, A., Hughes, T., Pas, J., Rychlewski, L., Travé, G., Aasland, R., Helmer-Citterich, M., Linding, R. & Gibson, T.J. (2010). ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic acids research*, **38**, D167–80.

Greene, L.H., Lewis, T.E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J.M. & Orengo, C.a. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic acids research*, **35**, D291–7.

Grünenfelder, B. & Winzeler, E.a. (2002). Treasures and traps in genome-wide data sets: case examples from yeast. *Nature reviews. Genetics*, **3**, 653–61.

Guerois, R., Nielsen, J.E. & Serrano, L. (2002). Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology*, **320**, 369–387.

Güldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W. & Stümpflen, V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic acids research*, **34**, D436–41.

Hernandez-Toro, J., Prieto, C. & De las Rivas, J. (2007). APID2NET: unified interactome graphic analyzer. *Bioinformatics* (*Oxford, England*), **23**, 2495–7.

Holm, L. & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics* (*Oxford, England*), **16**, 566–7.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J.a., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. & Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic acids research*, **37**, D211–5.

Huynen, M., Snel, B., Lathe III, W. & Bork, P. (2000). Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Research*, **10**, 1204–1210.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z. & Dunker, a.K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research*, **32**, 1037–49.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4569–74.

Janin, J. (2005). Assessing predictions of protein-protein interaction: the CAPRI experiment. *Protein science*, **14**, 278–283.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. & Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science (New York, N.Y.)*, **302**, 449–53.

Jayapandian, M., Chapman, A., Tarcea, V.G., Yu, C., Elkiss, A., Ianni, A., Liu, B., Nandi, A., Santos, C., Andrews, P., Athey, B., States, D. & Jagadish, H.V. (2007). Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic acids research*, **35**, D566–71.

Jefferson, E.R., Walsh, T.P., Roberts, T.J. & Barton, G.J. (2007). SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. *Nucleic acids research*, **35**, D580–9.

Jeong, H., Mason, S.P., Barabási, a.L. & Oltvai, Z.N. (2001). Lethality and centrality in protein networks. *Nature*, **411**, 41–2.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, a.a., Aflalo, C. & Vakser, I.a. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 2195–9.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. & Pandey, A. (2009). Human Protein Reference Database–2009 update. *Nucleic acids research*, **37**, D767–72.

Kiel, C., Foglierini, M., Kuemmerer, N., Beltrao, P. & Serrano, L. (2007). A genome-wide Ras-effector interaction network. *Journal of molecular biology*, **370**, 1020–32.

Kim, P.M., Lu, L.J., Xia, Y. & Gerstein, M.B. (2006a). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* (*New York, N.Y.*), **314**, 1938–41.

Kim, W.K., Henschel, A., Winter, C. & Schroeder, M. (2006b). The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS computational biology*, **2**, e124.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. & Greenblatt, J.F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–43.

Lehner, B. & Fraser, A.G. (2004). A first-draft human protein-interaction map. *Genome biology*, **5**, R63.

Letunic, I., Doerks, T. & Bork, P. (2009). SMART 6: recent updates and new developments. *Nucleic acids research*, **37**, D229–32.

Levy, E.D. (2007). PiQSi: protein quaternary structure investigation. *Structure* (*London, England : 1993*), **15**, 1364–7.

Mackay, J., Sunde, M., Lowry, J., Crossley, M. & Matthews, J. (2007). Protein interactions: is seeing believing? *Trends in biochemical sciences*, **32**, 530–531.

McDowall, M.D., Scott, M.S. & Barton, G.J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic acids research*, **37**, D651–6.

Miller, M.L., Jensen, L.J., Diella, F., J rgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S.a., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B.E., Yaffe, M.B.,

Brunak, S.r. & Linding, R. (2008). Linear motif atlas for phosphorylation-dependent signaling. *Science signaling*, **1**, ra2.

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, **247**, 536–40.

Neduva, V. & Russell, R.B. (2006). Peptides mediating interaction networks: new leads at last. *Current opinion in biotechnology*, **17**, 465–71.

Obenauer, J.C. (2003). Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, **31**, 3635–3641.

Ogmen, U., Keskin, O., Aytuna, a.S., Nussinov, R. & Gursoy, a. (2005). PRISM: protein interactions by structural matching. *Nucleic Acids Research*, **33**, W331–W336.

Orengo, C.A. & Thornton, J.M. (2005). Protein families and their evolution-a structural perspective. *Annual review of biochemistry*, **74**, 867–900.

Orengo, C.a., Michie, a.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure* (*London, England : 1993*), **5**, 1093–108.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.W., Ruepp, A. & Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics* (*Oxford, England*), **21**, 832–4.

Parrish, J.R., Gulyas, K.D. & Finley, R.L. (2006). Yeast two-hybrid contributions to interactome mapping. *Current opinion in biotechnology*, **17**, 387–93.

Pawson, T. & Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–52.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–8.

Ponting, C.P. & Russell, R.R. (2002). The natural history of protein domains. *Annual review of biophysics and biomolecular structure*, **31**, 45–71.

Reidel, B., Goldmann, T., Giessl, A. & Wolfrum, U. (2008). The translocation of signaling molecules in dark adapting mammalian rod photoreceptor cells is dependent on the cytoskeleton. *Cell motility and the cytoskeleton*, **65**, 785–800.

Rigaut, G., Shevchenko, a., Rutz, B., Wilm, M., Mann, M. & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, **17**, 1030–2.

Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D., Young, J., Yukich, B., Zardecki, C., Berman, H.M. & Bourne, P.E. (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research*, **39**, D392–401.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. & Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–8.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. (1998a). SMART, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5857–64.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. (1998b). SMART, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5857–64.

Schymkowitz, J.W.H., Rousseau, F., Martins, I.C., Ferkinghoff-Borg, J., Stricher, F. & Serrano, L. (2005). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 10147–52.

Seet, B.T., Dikic, I., Zhou, M.M. & Pawson, T. (2006). Reading protein modifications with interaction domains. *Nature reviews. Molecular cell biology*, **7**, 473–83.

Sigrist, C.J.a., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A. & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic acids research*, **38**, D161–6.

Sonnhammer, E.L., Eddy, S.R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–20.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K. & Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic acids research*, **39**, D698–704.

Stein, A., Panjkovich, A. & Aloy, P. (2009). 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic acids research*, **37**, D300–4.

Stein, A., Céol, A. & Aloy, P. (2011a). 3Did: Identification and Classification of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucleic acids research*, **39**, D718–23.

Stein, A., Mosca, R. & Aloy, P. (2011b). Three-dimensional modeling of protein interactions and complexes is going 'omics. *Current opinion in structural biology*, **21**, 200–8.

Stein, A., Mosca, R. & Aloy, P. (2011c). Three-dimensional modeling of protein interactions and complexes is going 'omics. *Current opinion in structural biology*, **21**, 200–8.

Stein, A., Rueda, M., Panjkovich, A., Orozco, M. & Aloy, P. (2011d). A Systematic Study of the Energetics Involved in Structural Changes upon Association and Connectivity in Protein Interaction Networks. *Structure* (*London, England : 1993*), **19**, 881–9.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C.,

Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. & Wanker, E.E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–68.

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. (2011a). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, **39**, D561–8.

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. (2011b). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, **39**, D561–8.

Toledo, F. & Wahl, G.M. (2006). Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nature reviews. Cancer*, **6**, 909–23.

Tuncbag, N., Kar, G., Gursoy, A., Keskin, O. & Nussinov, R. (2009). Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example. *Molecular BioSystems*, **5**, 1770–1778.

Tuncbag, N., Gursoy, A., Nussinov, R. & Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature Protocols*, **6**, 1341–1354.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.a., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, a., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J.M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–7.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

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.

Winter, C., Henschel, A., Kim, W.K. & Schroeder, M. (2006). SCOPPI: a structural classification of protein-protein interfaces. *Nucleic acids research*, **34**, D310–4.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T.P. & Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nature methods*, **6**, 75–7.

Xenarios, I., Salwinski, L., Joyce Duan, X., Higney, P., Kim, S.M. & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, **30**, 303–305.

Yu, H., Braun, P., Yildirim, M.a., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.F., Dricot, A., Vazquez, A., Murray, R.R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.S., Motyl, A., Hudson, M.E., Park, J., Xin, X., Cusick, M.E., Moore, T., Boone, C., Snyder, M., Roth, F.P., Barabási, A.L., Tavernier, J., Hill, D.E. & Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–10.

Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D., Swearingen, V., Yildirim, M.a., Yan, H., Dricot, A., Szeto, D., Lin, C., Hao, T., Fan, C., Milstein, S., Dupuy, D., Brasseur, R., Hill, D.E., Cusick, M.E. & Vidal, M. (2009). Edgetic perturbation models of human inherited disorders. *Molecular systems biology*, **5**, 321.

Zhu, H., Domingues, F.S., Sommer, I. & Lengauer, T. (2006). NOXclass: prediction of protein-protein interaction types. *BMC bioinformatics*, **7**, 27.