PATRICK ALOY

Structural bioinformatics and network biology

Proteins are the main perpetrators of most cellular tasks. However, they seldom act alone and most biological processes are carried out by macromolecular assemblies and regulated through a complex network of protein-protein interactions. Thus, modern molecular and cell biology no longer focus on single macromolecules but now look into complexes, pathways or even entire organisms. The many genome-sequencing initiatives have provided a near complete list of the components present in an organism, and post-genomic projects have aimed to catalogue the relationships between them. The emerging field of systems biology is now mainly centered on unravelling these relationships. However, all these interaction maps lack molecular details: they tell us who interacts with whom, but not how. A full understanding of how molecules interact can be attained only from high resolution three-dimensional (3D) structures, since these provide crucial atomic details about binding. These details allow a more rational design of experiments to disrupt an interaction and therefore to perturb any system in which the interaction is involved. Our main scientific interests are in the field of structural bioinformatics, in particular, the use of protein sequences and high-resolution 3D structures to reveal the molecular bases of how macromolecular complexes and cell networks operate.

Incorporating high-throughput proteomics experiments into structural biology pipelines

Recent years have seen the emergence of many large-scale proteomics initiatives that have identified thousands of new protein interactions and macromolecular assemblies. However, unfortunately, only a few of the complexes discovered meet the high-quality standards required to be promptly used in structural studies. This has thus created an increasing gap between the number of known protein interactions and complexes and those for which a high-resolution 3D structure is available. We have developed and validated a computational strategy to distinguish those complexes found in high-throughput affinity purification experiments that stand the best chances to be successfully expressed, purified and crystallised with little further intervention. Our study suggests that there are some 50 complexes recently discovered in yeast that could readily enter the structural biology pipelines. Indeed, we have used our target selection strategy to pick a list of 20 complex candidates whose structure determinations are going to be attempted by groups within 3D Repertoire, a large European integrated project that aims to solve the structures of all amenable protein complexes in yeast at the best possible resolution. The web version of the system is publicly available at http://targetselection.pcb.ub.es

Contextual specificity in peptide-mediated protein interactions

Protein interactions are central to virtually every major cellular function. While large protein-protein interfaces are typical in tightly associated macromolecular complexes, in most signalling events there is a globular domain in one protein that recognises a linear peptide from another, creating a relatively small interface. These interactions are predominantly found in regulatory networks and, due to their transient nature, are much more difficult to handle biochemically. Recently, large-scale experiments for the determination of peptide recognition profiles of interaction domains, and derivation of the corresponding patterns, have been developed, although transient peptide-mediated interactions are still underrepresented in high-throughput experiments. Even though binding is mediated by a small number of contacts formed by the residues in linear motifs, this type of interaction is extremely specific *in vivo*. For instance, it has been shown that the Pbs2 peptide is recognised only by the SH3 domain of Sho1 (its biological partner) and that it does not cross-react with any of the other 26 SH3 domains in yeast, although interactions with SH3 domains from other species are biophysically possible. More recently, another study has also shown that the binding specificity of PDZ domains is

Principal Investigator Patrick Aloy **Postdoctoral Fellows** Roberto Mosca, Andreas Zanzoni **PhD Students** Clara Berenguer, Marc Duocastella, Roland Pache, Amelie Stein **Research Assistant** Verónica Martínez **Lab Technician** Ricart Lluís

optimised across the 157 domains contained in the mouse proteome. However, bonds created between residues in linear motifs and globular domains, while sufficient to ensure binding, are too few to explain the high degree of specificity observed *in vivo*. It is thus, as occurs in phosphorylation events, the biological context that will ultimately determine the interaction specificity. This context has several aspects – certain subcellular localisation or expression patterns will determine whether proteins that are potential competitors for an interaction *in vitro* actually meet *in vivo* and thus evolve into niches of molecular recognition that allow them to bind only the desired target domain. Nevertheless, even within a cellular compartment several interaction domains and their complementary ligands are regularly expressed simultaneously, so yet more contextual information is required to achieve the observed specificity. This information is, to a great extent, contained in the residues surrounding the motif.

In the lab, we have systematically identified all instances of peptide-mediated protein interactions of known 3D structure and used them to study the individual contribution of motif and context to the global binding energy. We have found that, on average, the context is responsible for roughly 20% of the binding and plays a crucial role in determining interaction specificity, by either improving the affinity with the native partner or impeding non-native interactions. We have also examined and quantified the topological and energetic variability of interaction interfaces and have found a much higher heterogeneity in the context residues than in the consensus binding motifs (Figure 1). Our analysis partially revealed the molecular mecha-

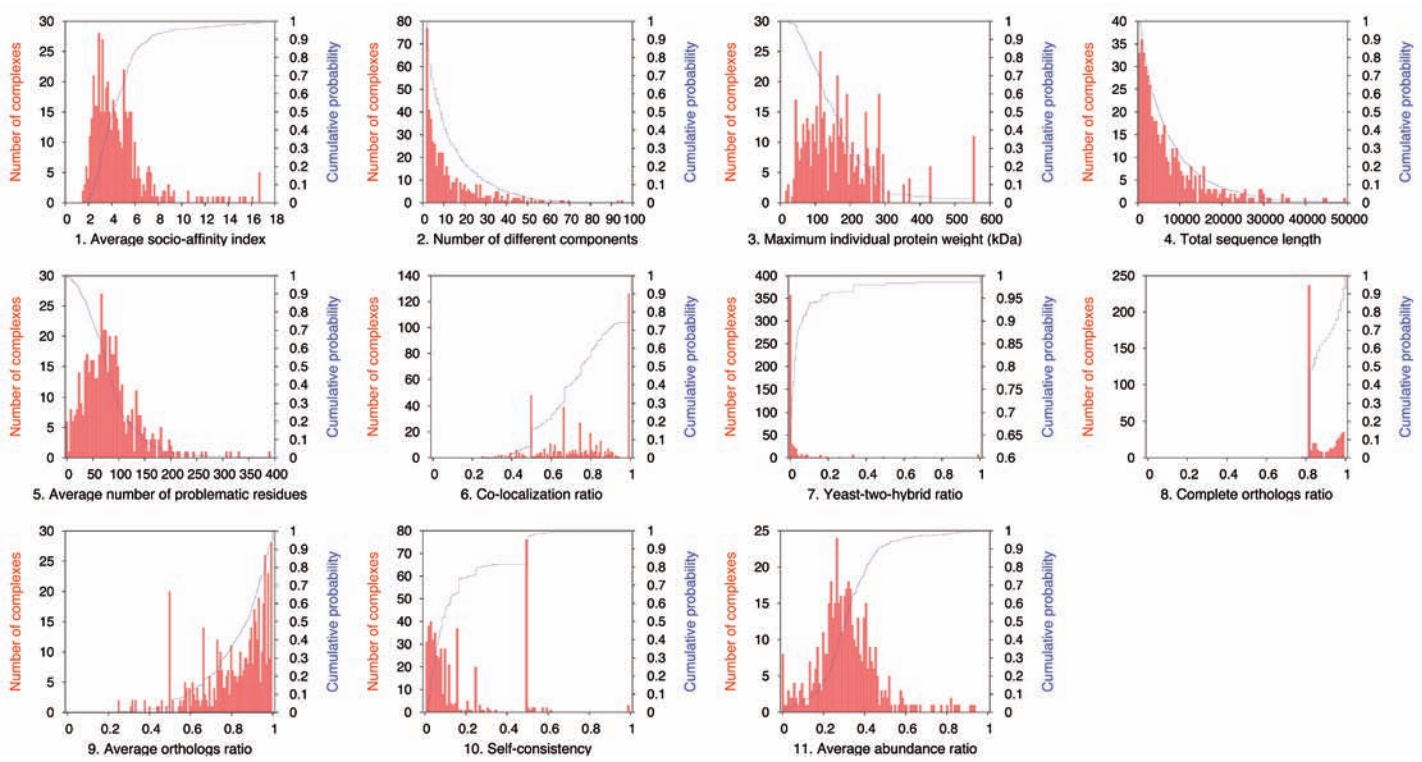

Figure 1. Distributions of partial scores and cumulative probabilities. *The distributions of partial scores for the eleven criteria used in the construction of the final feasibility score are shown in red, and the cumulative probabilities used for the normalisation in blue.*

nisms responsible for the dynamic nature of peptide-mediated interactions, and suggested a global evolutionary mechanism to maximise binding specificity. Finally, we have examined the viability of non-native interactions and highlighted cases of potential cross-reaction that might compensate for individual protein failures and establish backup circuits to increase the robustness of cell networks.

Exploiting gene deletion fitness effects to study the modular architecture of protein complexes under different growth conditions

An understanding of how individual genes contribute towards the fitness of an organism is a fundamental issue in biology. Although recent genome-wide screens have generated abundant data on quantitative fitness for single gene knock-outs, very few studies have systematically integrated other types of biological information to study how and why the deletion of specific genes gives rise to a particular fitness effect. In a recent study, we combined quantitative fitness data for single gene knock-outs in yeast with large-scale interaction discovery experiments to examine the effect of gene deletion on the modular architecture of protein complexes, under a range of growth conditions. Our analysis revealed that genes in complexes show more severe fitness effects upon deletion than other genes (Figure 2). However, in contrast to what has been observed in binary protein-protein interaction networks, we found that this was not related to the number of complexes in which they are present. We also observed that, in general, the modular components of protein complexes (*ie*, core and attachment proteins) are equally relevant for the complex machinery to function. However, when quantifying the importance of core and attachments in single complex variations, or isoforms, we found that this global trend originates from a combination of apparently unrelated factors, thereby indicating the presence of distinct fitness patterns in a single complex across growth conditions. Finally, our study also highlighted some interesting cases of potential functional compensation between protein paralogues and, perhaps, a new piece to fit in the histone-code puzzle.

Towards a molecular characterisation of Alzheimer's disease

In the last century, biomedical sciences were clearly immersed in a conceptual reductionism induced by the success of molecular biology. The development of methods to isolate and study individual cells and molecules has significantly increased our understanding of the nature of life and has led to considerable social advances, including the development of new medicines. Recent years have witnessed how the many genome sequencing projects have provided nearly complete lists of the macromolecules present in an organism, including humans. However, biological systems are often complex in nature, and the knowledge

of the components reveals relatively little about their function and organisation. The scientific community is now aware of the difficulties of predicting the behaviour of an intact organism from the individual actions of its parts in isolation and is rapidly moving to systems approaches, where global properties are also considered. In fact, most follow-up initiatives to the sequencing projects have been directed towards solving the systems' complexity and have focused on unveiling the millions of inter-relationships between macromolecules in an organism or monitoring how they co-ordinately change in response to a particular stimulus (*ie*, disease). Indeed, functional genomics initiatives are already delivering the first drafts of whole organism interactomes, gene expression profiles for many tissues and conditions, and the initial quantifications of metabolites in humans.

Pharmacological sciences have gone through a similar process, with traditional approaches being mostly reduced to the study, at the molecular level, of the target-compound duet. However, the truth is that phenotypic observations (*ie*, disease symptoms) are often the result of an incredibly complex combination of molecular events. This is because virtually every major biological process is not carried out by a single molecule but by large macromolecular assemblies and is often regulated through a complex network of transient interactions. Moreover, since most pathways are interconnected, slight changes in these transient regulatory networks can trigger one process or another, with completely different outcomes.

This reductionism has had striking consequences. For instance, many promising drug candidates have failed the last, and most expensive, sclinical phases because of the poorly understood mechanisms of action of the pathways they target or an inappropriate choice of the animal models, which proved ineffective at predicting off-target effects. It is thus clear that to increase the revenues of drug discovery, we need to improve our knowledge of the molecular mechanisms of disease by considering the full biological context of a drug target and moving beyond individual genes and proteins.

The main goal of our laboratory could be defined as the global molecular characterisation of pathological pathways through a combination of computational biology and interaction discovery techniques, in a real dry-wet cycle, where we use computational modelling to design the experiments needed to complement and complete the initial models (Figure 3). To this end, the recent creation of the Experimental Bioinformatics Lab, a joint initiative between IRB Barcelona and the Barcelona Supercomputing Center (BSC), has been crucial. The power of our approach is that we start from *in silico* modelling and therefore we are not restricted to the study of one or a few patho-physiological pathways. On the contrary, the first steps will involve a global modelling of all the human routes that might arise from known data, which will reveal novel and unexpected connections between them. We will then choose to further study those routes of most relevance from an academic or clinical perspective.

Starting from a set of seed proteins, an initial interactome is built using known protein-protein interactions. The resulting seed interactome or pathway is then extended and validated before putting its components into a spatio-temporal context based on gene expression data. Perturbation of the system finally allows us to unveil relationships between pathway topology and biological activity, with important implications for several kinds of clinical applications.

Indeed, we have already started to implement our approach to study the molecular bases of Alzheimer´s disease, where we have used three high-throughput interaction discovery approaches to screen over three thousand interactions. The initial phases of the project are yielding extremely interesting results (*ie,* we have discovered roughly 300 novel interactions and three potential new seed Alzheimer proteins), which will be pursued in the coming years.

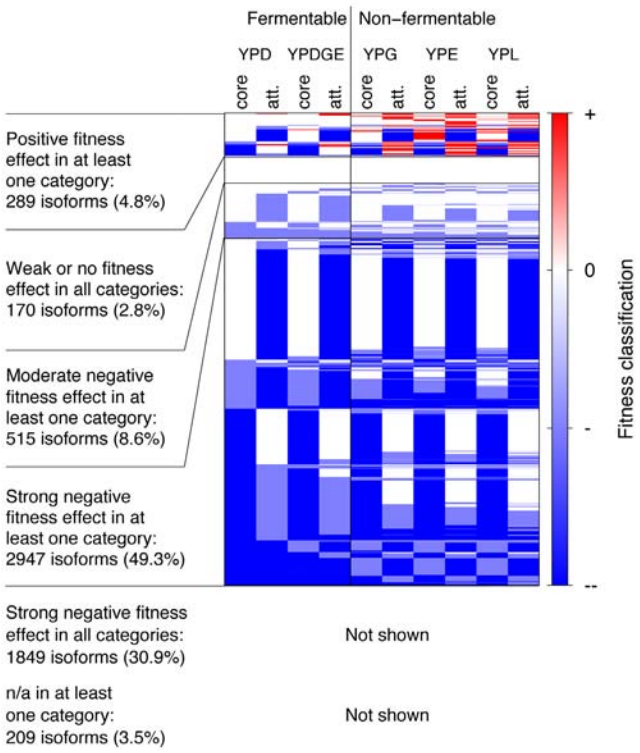

Figure 2. Fitness of the complex core and attachments of single isoforms across distinct growth conditions. *Expected fitness effects upon deletion of a random component of the given core or set of attachment proteins for all 5979 isoforms across the two fermentable and the three non-fermentable media are considered. The fitness values are partitioned into four categories: 'strong negative effect' (--/blue), 'moderate negative effect' (-/light-blue), weak or no effect' (0/white) and 'positive effect' (+/orange). Each line represents the fitness profile of a given isoform, treating the core and the attachments (att.) separately. 'n/a': the expected fitness effect is unknown due to a lack of quantitative fitness information for the genes in the respective core or attachments. When grouping the fitness profiles, we gave priority to n/a, positive, strong negative and moderate negative fitness effect in that order.*

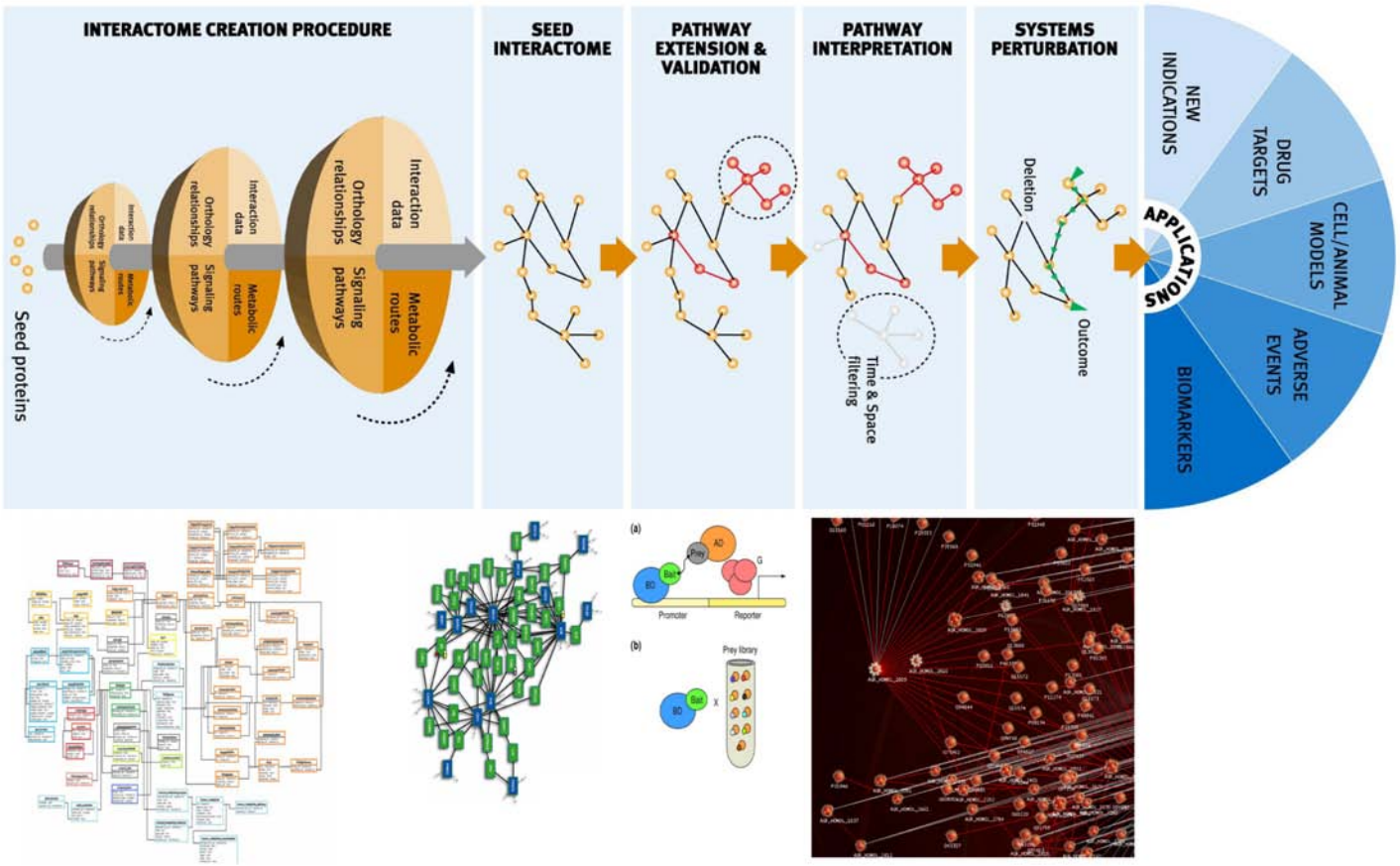

Figure 3. *Global strategy for the molecular characterisation of pathways and potential clinical applications.*

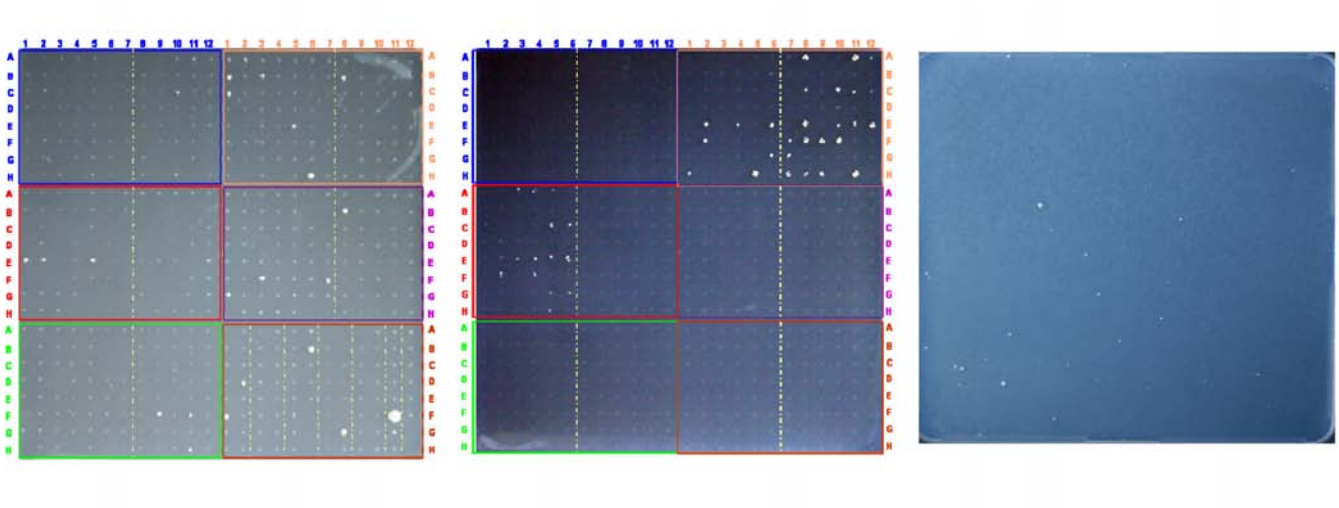

Figure 4. *Positive interactions found in the Alzheimer's interactome identified by co-transformation, mating and adult brain cDNA library two-hybrid screens.*

Publications

Pache RA and Aloy P. Incorporating high-throughput proteomics experiments into structural biology pipelines: identification of the low-hanging fruits. *Proteomics*, **8**(10), 1959-64 (2008)

Pache RA, Zanzoni A, Naval J, Mas JM and Aloy P. Towards a molecular characterisation of pathological pathways. *FEBS Lett*, **582**(8), 1259-65 (2008)

Parthasarathi L, Casey F, Stein A, Aloy P and Shields DC. Approved drug mimics of short peptide ligands from protein interaction motifs. *J Chem Inf Model*, **48**(10), 1943-48 (2008)

Russell RB and Aloy P. Targeting and tinkering with interaction networks. *Nat Chem Biol*, **4**(11), 666-73 (2008)

Stein A and Aloy P. A molecular interpretation of genetic interactions in yeast. *FEBS Lett*, **582**(8), 1245-50 (2008)

Stein A and Aloy P. Contextual specificity in peptide-mediated protein interactions. *PLoS ONE*, **3**(7), e2524 (2008)

Research networks and grants

A multidisciplinary approach to determine the structures of protein complexes in a model organism (3D-Repertoire)
European Commission, LSHG-CT-2005-512028 (2006-2009)
Principal investigator: Patrick Aloy

Aproximación bioinformática al estudio de la especificidad contextual en redes de interacciones entre proteínas y sus posibles aplicaciones biomédicas y biotecnológicas
Spanish Ministry of Science and Innovation, BIO2007-62426 (2007-2010)
Principal investigator: Patrick Aloy

Identificación de dianas secundarias y diseño de fármacos para enfermedades relacionadas con el envejecimiento mediante el análisis estructural y funcional de sus rutas biológicas
Spanish Ministry of Science and Innovation, PSE-010000-2007-1 (2007-2008)
Principal investigator: Patrick Aloy

Collaborations

Identification of genes regulator by FOXM1
Anastassis Perrakis, Nederlands Kanker Instituut (Amsterdam, The Netherlands)

Identification of potential phosphorylation targets for AURORA A kinase in human
Isabelle Vernos, Centre for Genomic Regulation—CRG (Barcelona, Spain)

Modular architecture of protein complexes and gene deletion fitness in yeast
Madan M Babu, MRC Laboratory of Molecular Biology (Cambridge, UK)

Network-based therapeutics
José Manuel Mas, Anaxomics Biotech (Barcelona, Spain)

New inhibitors of protein-protein interactions
Denis Shields, Trinity College (Dublin, Ireland)

Structural characterisation of molecular machines in yeast
Luís Serrano, Centre for Genomic Regulation (Barcelona, Spain)

Structural systems biology
Rob Russell, European Molecular Biology Laboratory (Heidelberg, Germany)