Structural bioinformatics/ Structural systems biology

Principal Investigator
Patrick Aloy (ICREA)

Postdoctoral Fellow
Andreas Zanzoni

PhD Students
Amelie Stein
Alejandro Panjkovich

MSc Student
Roland Pache

Patrick Aloy

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 centred mainly on unravelling these relationships. However, all these interaction maps lack molecular details: they tell us which molecules interacts with which, but not how. A full understanding of the way in which 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.

Target selection for complex structural genomics

Large-scale interaction discovery experiments are revealing the thousands of interactions and protein complexes responsible for most cellular functions, although they usually lack the molecular details that explain how the interactions occur. High-resolution 3D structures provide atomic information about the interaction interfaces. However, because of the difficulty of the experiments, the number of complexes of known structure is still limited. The launch of structural genomics initiatives focused on protein interactions and complexes could quickly fill the interaction space with structural details, thereby offering a new perspective on how cell networks operate at atomic level. Clear target selection strategies that rationally identify the key interactions and complexes that should be tackled first are fundamental to maximise return, minimise costs and prevent experimental difficulties. A complete interaction space filled with atomic-level details for each interaction, complex, signalling cascade and metabolic pathway would provide the perfect framework for future developments in systems biology.

We recently became a partner of the first-ever *complex* structural genomics initiative, named 3D repertoire, which seeks tos solve the structures of all amenable protein complexes in yeast at the highest resolution possible. The project involves 20 institutions across Europe and a multidisciplinary team of over 100 scientists in the fields of X-ray crystallography, NMR, electron microscopy and bioinformatics. We have been entrusted with the task of devising and implementing a strategy for selecting a set of complexes that stand a good chance of undergoing successful expression, purification and crystallisation. For this purpose, we have developed an automated procedure that combines several types of biological data (*eg*, socio-affinities, sub-cellular localisation, protein abundance, yeast two-hybrid experiments, *etc*) in a quantitative manner and scores each complex/sub-complex in yeast on the basis of its chances of success under standard expression and purification conditions. The system allows certain flexibility and we have also developed a web interface to permit users choose their own ranking criteria (http://gatealoy.pcb.ub.es/targetselection). The final result is a dynamic system to produce ranked lists of protein complexes/sub-complexes. The target selection strategy, the web-tool and the results have been disseminated among the partners of 3D repertoire and will shortly be available to the whole scientific community. We are currently expanding the target selection system to attach a confidence value to each complex/sub-complex and to extend it to other model organisms such as the fly, worm, mouse or human.

Contextual specificity in peptide-mediated protein interactions

Protein interactions are key to virtually every major process in a cell. While large protein-protein interfaces are typical in tightly associated macromolecular complexes (*eg,* RNA polymerase II), many transient interactions are mediated by a globular domain in one protein that recognises a small linear peptide in another (*eg,* SH3 domain recognising a proline-rich motif). However, although it has been shown that these motifs are enough to ensure binding, they are usually too short (4 to 11 residues) to achieve the high specificity observed in these interactions. It is thus a more general context outside the small linear peptide (*ie,* other interacting residues, subcellular localisation, expression patterns, *etc*) that will ultimately determine the interaction between two given proteins. (See Figure 1.)

In the lab, we have used high-resolution 3D structures of interacting protein pairs to explore the contribution of the binding motif and surrounding residues (the context) in all known domain-ligand interaction types. We have found that, on average, contextual contacts account for roughly 30% of the binding energy. We have also seen that the central motif itself is fairly unspecific as it has the capacity to bind to many homologous domains. Therefore it is the context, to a large extent, that is responsible for the high affinity shown by this type of interaction, either by improving the binding energy with the native partner or preventing non-native interactions. We are currently exploring the bearings that our findings have in systems biology, since they might provide a rationale for several compensatory effects observed in knock-out networks and phenotypic profiles, and in synthetic biology, where specificity information is instrumental in the construction of artificial cellular circuits. (See Figure 2.)

Pushing structural details into protein interaction networks

Although there is a growing gap between the number of interactions detected and those for which the 3D structure is known, it is equally true that, since the mid-nineties, crystallographers have been solving structures at a rate of over one thousand interactions every year. Thus, the Protein Data Bank currently contains many thousands of interactions for which structural data are available, which implies that it is increasingly possible to model structures for protein interactions on the basis of those observed previously. Like most modelling efforts, accuracy depends greatly on the degree of sequence identity between the target and the template onto which it is modelled. When modelling an interaction, the choice of the template is all the more crucial because the use of the wrong template may lead to protein interactions through the incorrect interface. This is roughly analogous to modelling a single protein on another that has a different fold. Encouragingly, however, when sequence similarity is high (for example, >25-30% identity) proteins are likely to interact in the same way, although exceptions are possible. There are instances where interactions are structurally similar despite there being no sequence similarity; the trick is to find them. The last five years have seen the emergence of a new class of techniques that use protein-protein complexes for which coordinate data are available to model interactions between their homo-

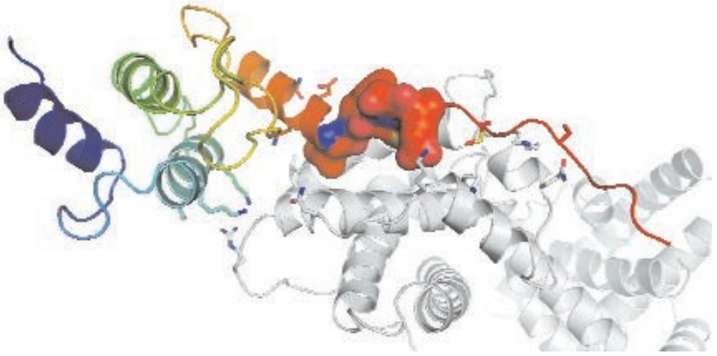


Figure 1. *Example of contextual specificity. Interaction between the retinoblastoma protein (grey) and a viral peptide (rainbow). The consensus motif ([LI].C.[DE]) responsible for the binding is shown in a surface representation. Selected contextual interactions are highlighted in a stick model, some of which are at a considerable distance from the motif.*

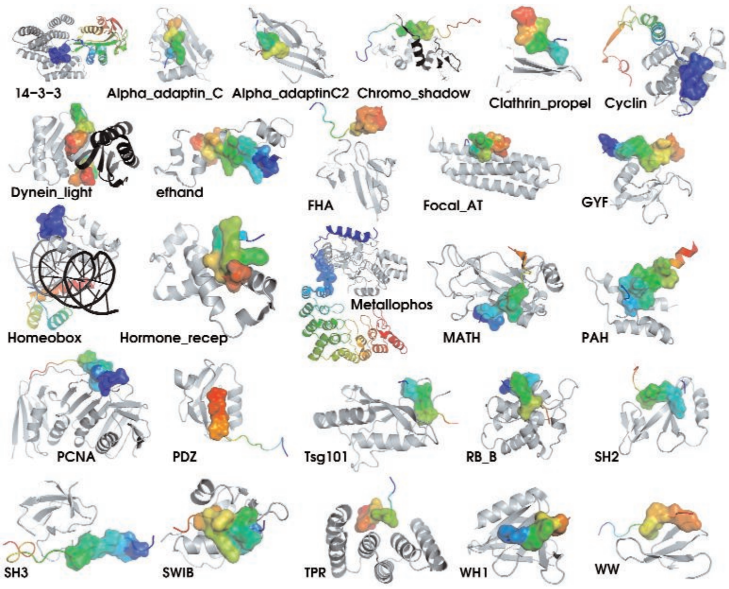


Figure 2. *Transient domain-peptide interactions of known 3D structure. Collection of representative structures of interaction domains (grey) bound to their partner proteins (rainbow) containing the binding linear motifs (surface).*

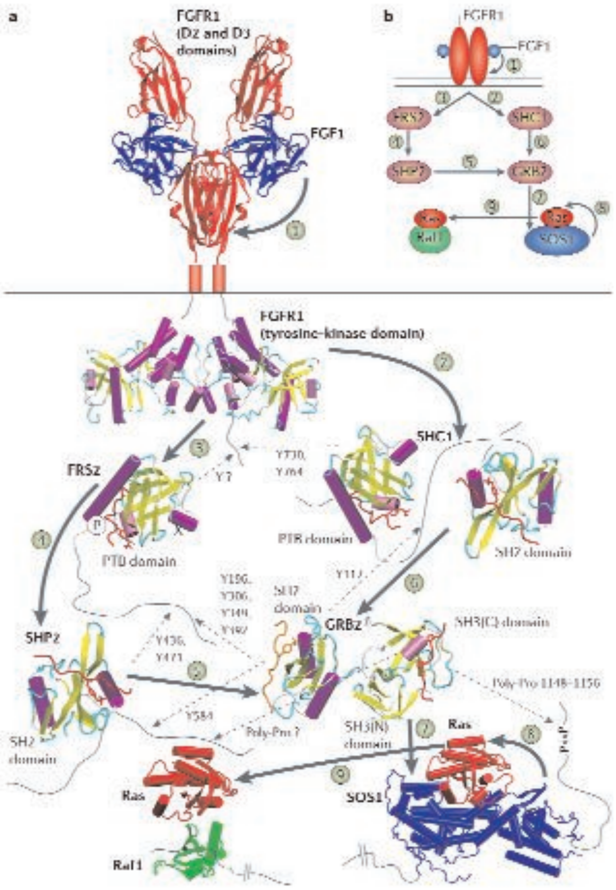


Figure 3. The fibroblast growth factor signalling pathway. Structural details inferred for the fibroblast growth factor (FGF) pathway (a) as a means to complement its classical blob-representation shown in the text books (b). X-ray, NMR or modelled structures are shown in diagram format (α-helices are shown as cylinders and β-strands are depicted as arrows). Structures of complexes of two large proteins (for example, FGF/FGFR) are coloured according to chain (that is, each separate polypeptide is a different colour); structures involved in domain-peptide or phosphorylation interactions are coloured according to the secondary structure of the domain (helices are magenta and β-strands yellow) and the interacting peptides are red, or orange or are shown schematically. Black arrows denote the activation events highlighted in the schematic diagram, dashed arrows show interactions between domains in one protein and particular regions on another; the labels on these arrows indicate the residues where the domain binds (when known). (Aloy and Russell, 2006.)

logues. However, these approaches are far from perfect, and they suffer when the interactions involve conformational changes at the interface, or when the modelled interfaces contain insertions or deletions, with respect to the template, that are not accurately modelled. Moreover, they are usually unable to sort out the correct specificity between members of two interacting families and the results given often do not have any biophysical meaning (*ie*, there is no correlation between computationally derived scores and, for instance, dissociation constants). We are now working on a novel approach to tackle the specificity problem and to predict binding energies for domain-domain interactions by means of empirical pair potentials. We are also developing an automated strategy to use these potentials to predict new protein-protein interactions on a genome scale. Being structure-based, the resulting networks will not only tell us which molecules interacts with which but will also give quasi-atomic information as to how these interactions occur. This level of detail will permit a much more rational design of experiments to disrupt an interaction and therefore to perturb the global behaviour of the network.

Structural systems biology: explaining macroscopic effects at molecular level

One of the major goals of the emerging field of systems bology is to explain macroscopic effects at molecular level. This is to build computational models to simulate the behaviour of complex systems, such as limb development in mice, or the cellular circuits responsible for certain phenotypes in yeast or *Arabidopsis*. We aim to combine the knowledge we have acquired from the above projects (*ie*, the molecular determinants of specificity in the different types of protein-protein interactions, structure-based interaction networks, *etc*) with other types of data available (*ie*, protein expression, interaction kinetics, *etc*) in order to propose potential new pathways that explain the observed phenotypic profiles observed in model organisms. One of the first steps that we have taken towards this end is to push quasi-atomic details into known signalling pathways, wherever possible, to convert the classical, and not very informative, blob-diagrams into something more structurally meaningful that can provide much clearer insights into the events that occur. (See Figure 3.)

PUBLICATIONS

Aloy P and Russell RB (2006) Structural systems biology: modeling protein interaction networks. Nature Rev Mol Cell Biol, 7:188-197

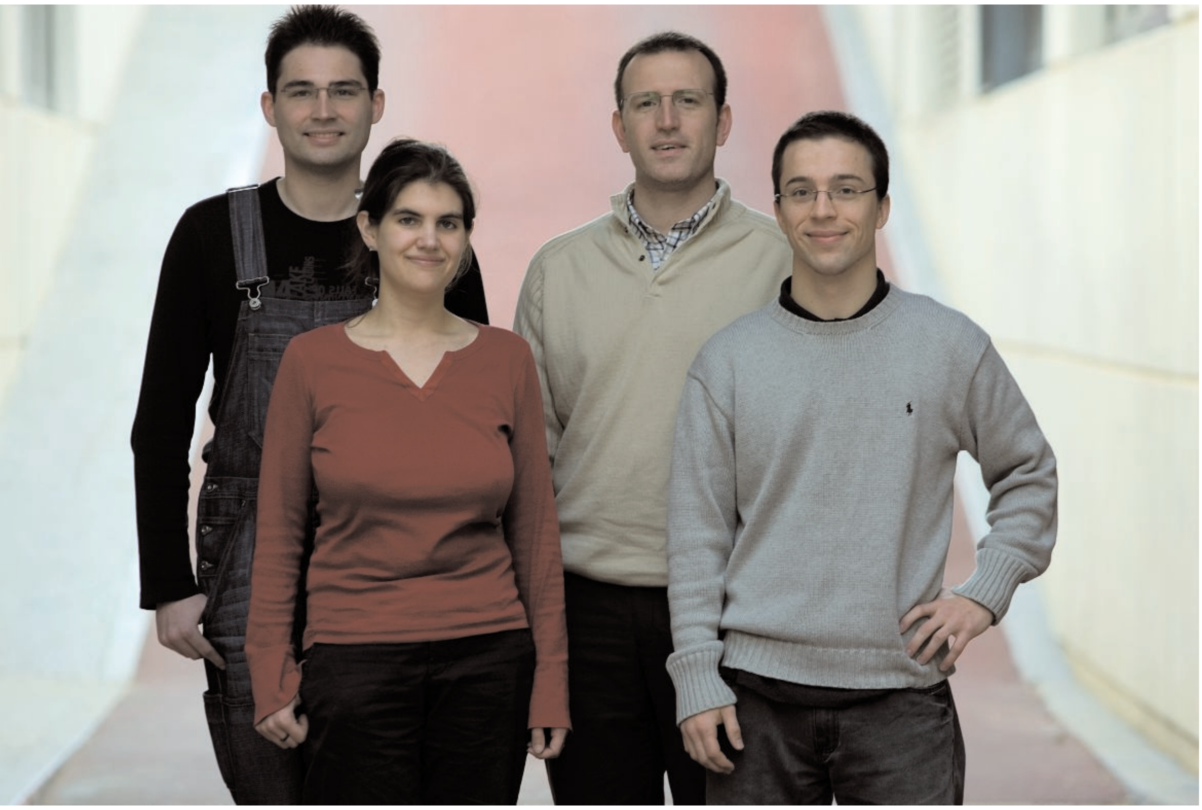
Bravo J and Aloy P (2006) Target selection for complex structural genomics. Curr Opin Struct Biol, 16:385-392

Dopazo J and Aloy P (2006) Discovery and hypothesis generation through bioinformatics. Genome Biol, 7:307

Gavin AC*, Aloy P*, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB and Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature, 440:631-636

RESEARCH NETWORKS AND GRANTS

A multidisciplinary approach to determine the structures of protein complexes in a model organism
European Integrated Project, European Commission: 2005
Project Coordinator: Luis Serrano



Patrick Aloy's group, March 2006.