

# Structural Computational Biology Group

## Summary

The main interest of our Group is to understand the organisation and evolution of gene/protein networks and in particular the relationship between protein/gene specific interactions with cancer related processes.

Our research centres on the problem of functional specificity and selective interactions between molecular components. On the technical level we focus on the development of computational methods for the integration of information from heterogeneous genomic resources. Research is developed in the context of cancer genome projects.

## Strategic Goals

- Analyse the function and structure of proteins related to cancer through the use of molecular networks to integrate molecular, genomic and phenotypic data
- Develop new methods and software platforms for the integration and representation of cancer genomic data, including the statistical analysis of CGH arrays – with particular emphasis on those derived from new sequencing technologies
- Apply bioinformatics technology for the integration of complex data sets in collaboration with large networks (NIH ENCODE) and the International Cancer Genome Consortium

## Alfonso Valencia *Group Leader*

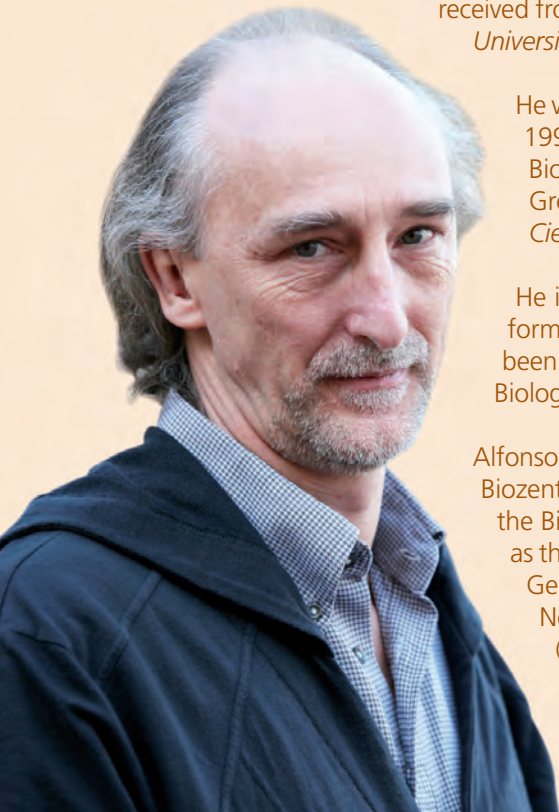
Alfonso Valencia is a biologist with formal training in population genetics and biophysics which he received from the *Universidad Complutense de Madrid*. He was awarded his PhD in 1988 at the *Universidad Autónoma de Madrid*.

He was a Visiting Scientist at the American Red Cross Laboratory in 1987 and from 1989 – 1994 was a Postdoctoral Fellow at the laboratory of C. Sander at the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. In 1994 he set up the Protein Design Group at the *Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC)* in Madrid where he was appointed as Research Professor in 2005.

He is a Member of the European Molecular Biology Organisation (EMBO), Founder and former Vice President of the International Society for Computational Biology where he has been Chair of the Systems Biology and/or Text Mining Tracks of the main Computational Biology Annual Conference (ISMB) since 2003.

Alfonso Valencia serves on the Scientific Advisory Board of the Swiss Institute for Bioinformatics, Biozentrum, Basel; the INTERPRO database; the Spanish Grant Evaluation Agency (ANEP); the Biotechnology and Biological Sciences Research Council (BBSRC) expert panel; as well as the Steering Committee of the European Science Foundation Programme on Functional Genomics (2006 – 2011). His Group participates in the three main Bioinformatics Networks of Excellence organised under the 6th European Framework Programme (BioSapiens, EMBRACE and ENFIN).

Alfonso Valencia is Co-Executive Editor of *Bioinformatics*, serves on the Editorial Board of *FEBS J*, *EMBO Journal* and *EMBO Reports*, among others. He is Director of the Spanish National Bioinformatics Institute (INB).





**Staff scientists:** Ildefonso Cases (until June), Ramón Díaz, Gloria Fuentes (until June), Ana M. Rojas (until June), Michael Tress. **Post-doctoral fellows:** Anäis Baudot, Milana Morgenstern (since September), Daniel Rico, Ashish V. Tendulkar (july through September), Jan J. Wesselink (until April). **Graduate students:** César Boullosa, Iakes Ezcurdia, José M. González-Izarzugaza, Gonzalo López, Paolo Maietta (since May), Antonio Rausell. **Technicians:** Ángela del Pozo, Martín Krallinger, David A. Juan.

## Highlights

### Protein structure and function

We have continued to develop our Firestar and FireDB methods for the prediction of protein binding sites as well as their utilisation for the evaluation of prediction methods in the context of “Critical Assessment of techniques for protein Structure Prediction” (CASP) (see Proteins CASP09 Special issue). Our work in the CASP evaluation of prediction methods opens new avenues for the integration of methods such as the recent demonstration of the potential of contact prediction methods to screen and validate structural models.

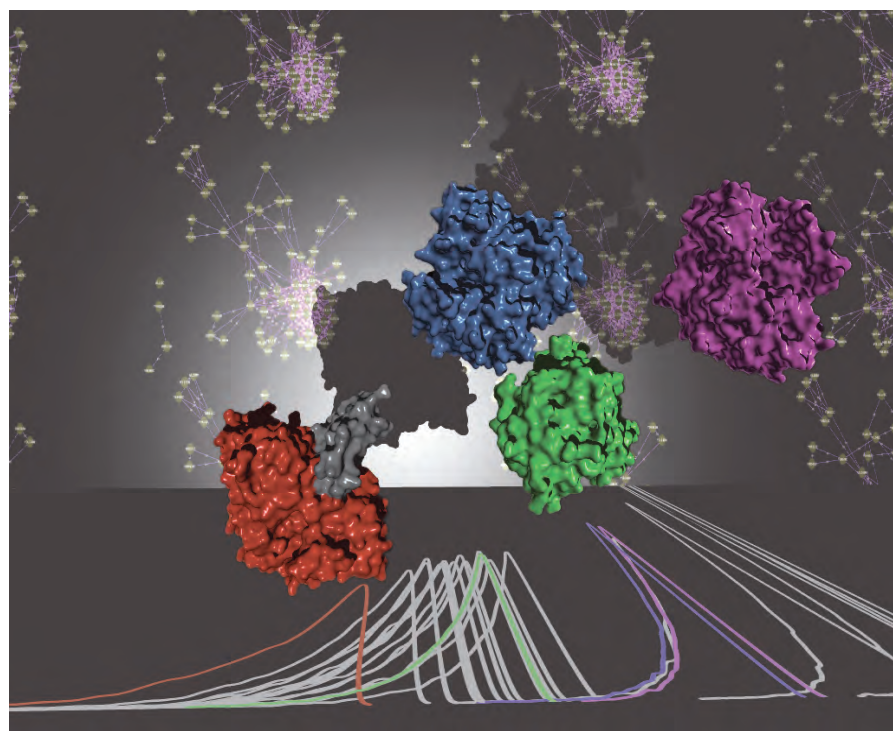
Combining the results from those prediction methods, the available protein structures and published experimental data on mutations and interactions, we have proposed new models for the specific interactions of ras-p21 and its main effectors.

On the methodological side we have developed a new high throughput approach for the systematic prediction of protein interactions based on the physical characteristics of the surfaces of interacting proteins (Figure 1).

### Biological text mining

In the field of biological text-mining we organised the BC II.5 BioCreative Challenge – a community-wide effort to evaluate information extraction systems applied to biological problems. BCII.5 is dedicated to assisting authors

generate Structured Digital Abstracts through the application of text-mining methodology, in accordance with the model implemented by FEBS letters in collaboration with the Molecular INTERaction database (MINT). This initiative will be followed by discussions with publishers and databases for the practical integration of text-mining methodology to link the information highlighted by the authors on their own papers directly with the



**Figure 1:** Artistic structural representation of interacting proteins and their relation in known protein complexes, represented together with profiles of their interaction scores predicted using a new docking-based approach. This new approach is able to predict interaction partners based on the structure of the corresponding unbound proteins.

corresponding database information. BC III will be held in Washington DC in September 2010.

During 2009 we have developed specific applications in text-mining technology for the prediction and classification of proteins involved with spindle formation, cell-cycle control, and chromosome condensation. These developments are a result of participating in the Experimental Network for Functional INtegration (ENFIN) Network of Excellence.

The fusion of our efforts in protein structure analysis and text-mining have resulted in the analysis of the distribution of cancer-related mutations in protein kinases, a work partially carried out in collaboration with C. Orengo's Group at the University College of London (UCL), UK. We have been able to duplicate the number of mutations characterised in databases (including the artificially introduced ones) and natural variants by extracting them directly from the original references (Figure 2).

## Protein complexes and networks

We have continued to work on protein interaction networks with the aim of using molecular networks as the framework for analysing cancer genome data, serving as links between molecular/genomic cancer data and clinical/phenotypic information. In the context of the Innovative Medicines Initiative (IMI) initiative in toxicogenomics (e-TOX) we will be further developing this idea; combining genomics and phenotypic (toxicology) information by applying bioinformatics, systems biology and text-mining technology.

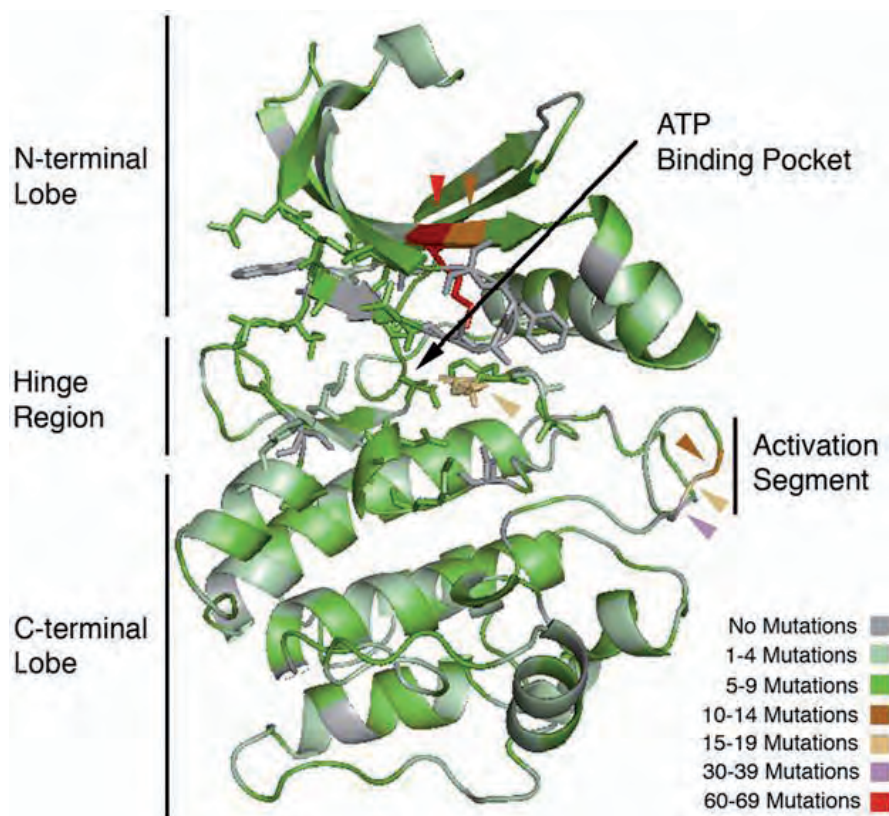
As a first incarnation of this principle we have carried out a systematic comparison of cancer genome data associated to specific tissues, using the similarity of the distribution of related genes in the network of known protein interactions as metric. As expected, genes and pathways commonly implicated in cancer are clearly detected, but interesting novel associations linking cancer types and specific pathways that accumulate a significant number of mutated genes are also highlighted.

## Human variation data and cancer

The Group is involved in the analysis of a variety of cancer genome data, including various platforms and systems (SNP arrays, GCH arrays, expression data, exon sequencing, and DNA methylation data), as well as Next Generation Sequence data (chip-seq and exon sequencing).

We are also developing the infrastructure analysing the results of the Spanish initiative in the context of the International Cancer Genome Consortium.

We have developed new methods for the statistical analysis of CGH-arrays and for the selection of candidate genes/functions from high-throughput genomic information. We are particularly interested in the study of genetic interactions using our basic ideas on gene/protein interactions and networks.



**Figure 2:** Distribution of mutations in protein kinases extracted from the literature and represented in the canonical protein kinase structure. The accumulation of mutations in key functional areas is clearly visible.

# Publications

Fuentes G, Valencia A (2009). Ras classical effectors: new tales from *in silico* complexes. *Trends Biochem Sci* 34, 533-539.

Sequeira-Mendes J, Díaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gómez M (2009). Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet* 5, e1000446.

Babel I, Barderas R, Díaz-Uriarte R, Martínez-Torrecedrera JL, Sánchez-Carbayo M, Casal JI (2009). Identification of tumor-associated autoantigens for the diagnosis of colorectal cancer in serum using high density protein microarrays. *Mol Cell Proteomics* 8, 2382-2395.

Trigo A, Valencia A, Cases I (2009). Systemic approaches to biodegradation. *FEMS Microbiol Rev* 33, 98-108.

Baudot A, Real FX, Izarzugaza JM, Valencia A (2009). From cancer genomes to cancer models: bridging the gaps. *EMBO Rep* 10, 359-366.

Carbajosa G, Trigo A, Valencia A, Cases I (2009). Bionemo: molecular information on biodegradation metabolism. *Nucleic Acids Res* 37, D598-D602.

Andres Leon E, Ezkurdia I, García B, Valencia A, Juan D (2009). EclD. A database for the inference of functional interactions in *E. coli*. *Nucleic Acids Res* 37, D629-D635.

Krallinger M, Rodriguez-Penagos C, Tendulkar A, Valencia, A (2009). PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction. *Nucleic Acids Res* 37, W160-W165.

Morrissey ER, Diaz-Uriarte R (2009). Pomelo II: finding differentially expressed genes. *Nucleic Acids Res* 37, W581-W586.

Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML (2009). Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform* 10, 233-246.

Rueda OM, Diaz-Uriarte R (2009). RJaCGH: Bayesian analysis of aCGH arrays for detecting copy number changes and recurrent regions. *Bioinformatics* 25, 1959-1960.

Fuentes G, Oyarzabal J, Rojas AM (2009). Databases of protein-protein interactions and their use in drug discovery. *Curr Opin Drug Discov Devel* 12, 358-366.

Fernandez-Ballester G, Beltrao P, Gonzalez JM, Song YH, Wilmanns M, Valencia A, Serrano L (2009). Structure-based prediction of the *Saccharomyces cerevisiae* SH3-ligand interactions. *J Mol Biol* 388, 902-916.

Bacardit J, Stout M, Hirst JD, Valencia A, Smith RE, Krasnogor N (2009). Automated alphabet reduction for protein datasets. *BMC Bioinformatics* 10, 6.

Izarzugaza JM, Baresic A, McMillan LE, Yeats C, Clegg AB, Orengo CA, Martin AC, Valencia A (2009). An integrated approach to the interpretation of single amino acid polymorphisms within the framework of CATH and Gene3D. *BMC Bioinformatics* 10, S5.

Krallinger M, Izarzugaza JM, Rodriguez-Penagos C, Valencia A (2009). Extraction of human kinase mutations from literature, databases and genotyping studies. *BMC Bioinformatics* 10, S1.

Rueda OM, Diaz-Uriarte R (2009). Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously. *BMC Bioinformatics* 10, 308.

Tress ML, Ezkurdia I, Richardson JS (2009). Target domain definition and classification in CASP8. *Proteins* 77, 10-17.

Izarzugaza JM, Redfern OC, Orengo CA, Valencia A (2009). Cancer-associated mutations are preferentially distributed in protein kinase functional sites. *Proteins* 77, 892-903.

López G, Ezkurdia I, Tress ML. (2009). Assessment of ligand binding residue predictions in CASP8. *Proteins* 77, 138-146.

Ezkurdia I, Graña O, Izarzugaza JM, Tress ML (2009). Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins* 77, 196-209.

Krallinger M, Rojas AM, Valencia A (2009). Creating reference datasets for systems biology applications using text mining. *Ann NY Acad Sci* 1158, 14-28.

Baudot A, Gómez-López G, Valencia A (2009). Translational disease interpretation with molecular networks. *Genome Biol* 10, 221.

Juncker AS, Jensen LJ, Pierleoni A, Bernsel A, Tress ML, Bork P, von Heijne G, Valencia A, Ouzounis CA, Casadio R, Brunak S (2009). Sequence-based feature prediction and annotation of proteins. *Genome Biol* 10, 206.

## Book chapter

Tress M, Bujnicki JM, López G, Valencia A (2009). Integrating prediction of structure, function and interactions. p.259-280. In: Prediction of Protein Structures and Interactions, Janusz Bujnicki (ed.), Wiley, John & Sons.

## Awards and Recognition

Expert Member, BBSRC Strategy Research Committee (2009-2012), UK

Member, ERC Advanced Grant panel (2009-2011)

Member, Spanish Grant Evaluation Agency Expert Panel (2009-2011)