

# Structural Computational Biology Group

48

Scientific Report 2010 *crío*



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 Alfonso Valencia 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. He was honoured as ISCB-Fellow in 2010.

Alfonso Valencia serves on the Scientific Advisory Board of the European Molecular Biology Laboratory; the Swiss Institute for Bioinformatics, Biozentrum, Basel; the INTERPRO database; the Spanish Grant Evaluation Agency (ANEP); as well as the Steering Committee of the European Science Foundation Programme on Functional Genomics (2006-2011).

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

## Summary

The main interest of our group is to understand the organisation and evolution of gene/protein networks particularly the relation between protein/gene specific interactions with cancer related processes.

Our research centres on the problem of functional specificity and selective interactions between molecular components. We develop computational methods for the integration of data from heterogeneous genomic resources in the context of cancer genome research.

## Strategic Goals

- Analyse the function and structure of proteins related to cancer with emphasis on specific molecular interactions
- Develop new methods and software platforms for the extraction, integration and representation of cancer genomic data, including the statistical analysis of molecular, genomic and phenotypic information, with particular attention to those deriving from new sequencing technologies in collaboration with cancer genome projects
- Design the next generation of computational methods for the interpretation of personalised cancer genomic information





**Staff scientists:** Ramón Díaz and Michael Tress. **Post-doctoral fellows:** Anais Baudot (until July), Milana Morgenstern, Daniel Rico, Miguel Vázquez (since August) and Jorge A. Zamora (since November). **Graduate students:** César Boullosa, Iakes Ezcurdia, José M. González-Izarzugaza, Cristina Ibáñez (since October), Florian Leitner, Gonzalo López (until September), Paolo Maietta and Antonio Rausell. **Technicians:** Ángela del Pozo, David A. Juan and Martin Krallinger.

## Highlights

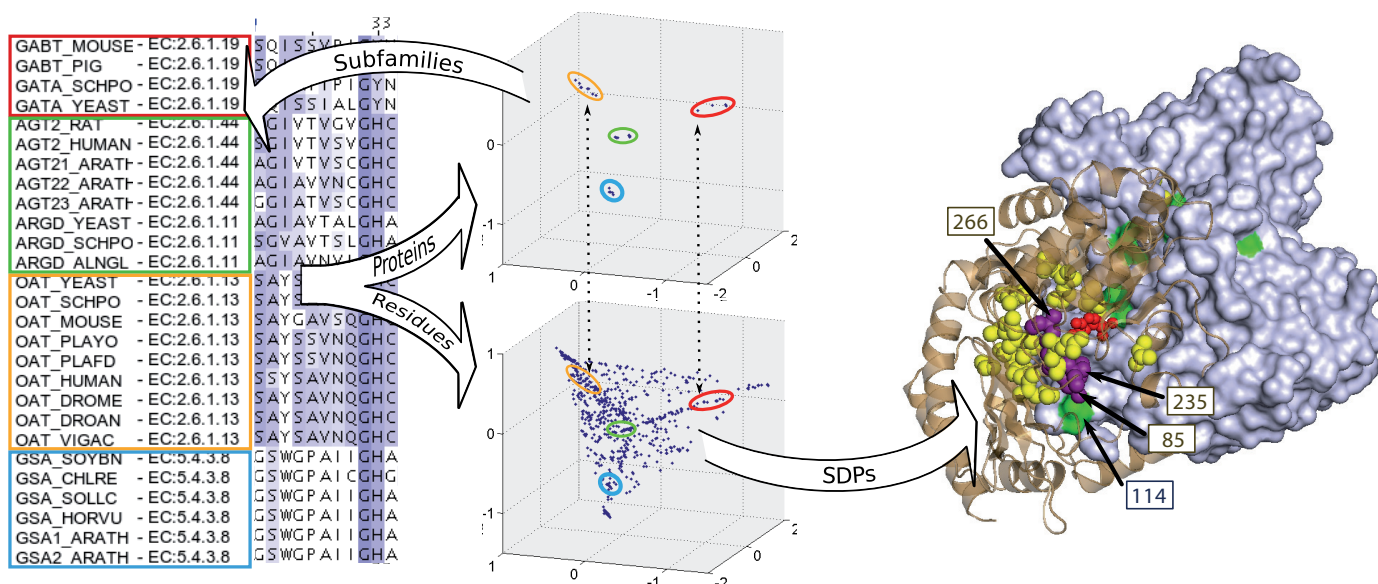
### Protein structure and function

We continue to develop the Firestar/FireDB method for predicting protein binding sites based on the extrapolation of information from known protein structures and modelling approaches. We are complementing this information by developing *ab initio* methods for the prediction of sites responsible for specific protein interactions with their substrates, cofactors and protein partners (Rausell A. et al., *PNAS* 2010). The results from these combined predictions will be assessed during the meeting Critical Assessment of techniques for protein Structure Prediction (CASP), Asilomar CA (USA), December 2010.

The study of binding sites represents one source of data that has been incorporated to evaluate the potential consequences of alternative splicing at the protein level. We use the complete APRIS system for the annotation of the human genome for highlighting potential functional isoforms – in collaboration with the HAVANA annotation team at the Sanger Institute (UK) – and to study trans-splicing in the context of the ENDODE and Consolider E-science projects (Figure 1).

### Biological text mining

In the field of biological text mining we co-organised the BioCreative III



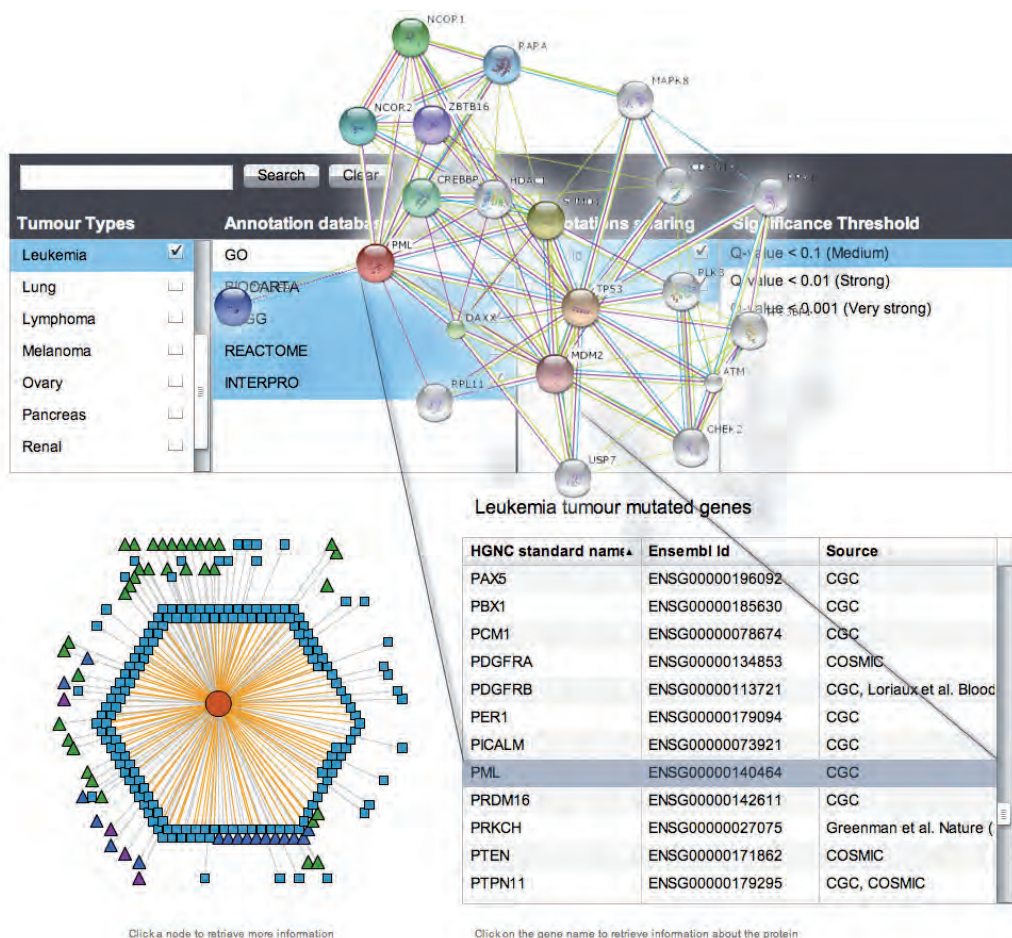
**Figure 1:** Simultaneous detection of protein subfamilies and associated residues. The example depicts the structure of the class III aminotransferase family showing the binding site (in red spheres) and predicted residues highlighted in a yellow/violet spacefill and with a green surface (taken from Rausell A. et al., *PNAS* 2010).

Challenge – a community-wide effort to evaluate information extraction systems applied to biological problems. BCIII assists experimental biologists and database annotators organise and retrieve information from manuscripts through the application of text mining technology. The results of the assessment of current methods as well as the discussions on text-mining technology and standards for textual data repositories during the BioCreative meeting in Bethesda MD (USA), September 2010, will be published in 2011 in a special issue of *BMC Bioinformatics*. These publications will complement those of BioCreative II.5, which were published earlier this year (Leitner F. et al., *Nat Biotechnol* 2010, *FEBS Lett* 2010).

We are currently using text-mining technologies to advance a number of fields that previously required analysis of the functions of large sets of proteins/genes derived from high-throughput

data. We developed two projects through the Experimental Network for Functional Integration (ENFIN) Network of Excellence collaboration on the composition analysis of the human spindle complex and the genes potentially implicated in chromosome condensation. The application of this methodology is an essential tool to discover potential functions and molecular interactions between candidate genes that may be associated with certain types of cancer such as melanoma, and in particular for our work in the Innovative Medicines Initiative (IMI) in toxicogenomics (e-TOX) and the new ASSET EU project "Assessing Sensitivity of Embryonal Tumours".

Our combined efforts in protein structure/function analysis and text mining have proven valuable in the analysis of the distribution of cancer-related mutations in protein kinases – work partially carried out in collaboration with C. Orengo's group at the University College London (UK).



**Figure 2:** Visualisation of the altered functions (pathways) common to various cancer types, obtained from available cancer genome studies (Baudot A. et al., *EMBO Rep* 2010). The system is available at <http://contexts.bioinfo.cnio.es/cancer-processes/tool>.

## Protein complexes and cancer networks

We have continued to promote the use of molecular networks as the framework for analysing cancer genome data, serving as potential links between molecular/genomic cancer data and clinical/phenotypic information.

We have carried out a first systematic comparison of proteins and functions associated to generic cancer types using genomic information extracted from databases (Baudot A. et al., *EMBO Rep* 2010). We are currently working on the refinement and extension of this approach by developing new methodologies for the redefinition of molecular pathways. This redefinition will increase the functional space available for the interpretation of these types of large data sets. We are also developing a more refined definition of cancer types which includes individual cancer data, as well as a more

comprehensive definition of mutations by including data from various platforms and systems (SNP arrays, GCH arrays, and Next Generation Sequence data on chip-seq, exon sequencing, and DNA methylation data). For this work, we are also consolidating the infrastructure for handling and analysing the results of the Spanish initiative in the context of the International Cancer Genome Consortium (ICGC publication).

On the methodological side, we have continued to work on the statistical analysis of NGS and CGH-arrays, including the evaluation of current methods for studying gene/protein interactions (epistasis) and their extension using concepts such as protein interaction and gene control networks.

## Publications

International Cancer Genome Consortium. (2010). International network of cancer genome projects. *Nature* 464, 993-998.

Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, Licata L, Hirschman L, Cesareni G, Valencia A (2010). The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. *Nat Biotechnol* 28, 897-899.

Rodríguez-Santiago B, Malats N, Rothman N, Armengol L, Garcia-Closas M, Kogevinas M, Villa O, Hutchinson A, Earl J, Marenne G, Jacobs K, Rico D, Tardón A, Carrato A, Thomas G, Valencia A, Silverman D, Real FX, Chanock SJ, Pérez-Jurado LA (2010). Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *Am J Hum Genet* 87, 129-138.

Rausell A, Juan D, Pazos F, Valencia A (2010). Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci USA* 107, 1995-2000.

Baudot A, de la Torre V, Valencia A (2010). Mutated genes, pathways and processes in tumours. *EMBO Rep* 11, 805-810.

Carilla-Latorre S, Gallardo ME, Annesley SJ, Calvo-Garrido J, Graña O, Accari SL, Smith PK, Valencia A, Garesse R, Fisher PR, Escalante R (2010). MidA is a putative methyltransferase that is required for mitochondrial complex I function. *J Cell Sci* 123, 1674-1683.

Glaab E, Baudot A, Krasnogor N, Valencia A (2010). TopoGSA: network topological gene set analysis. *Bioinformatics* 26, 1271-1272.

Tendulkar AV, Krallinger M, de la Torre V, López G, Wangikar PP, Valencia A (2010). FragKB: structural and literature annotation resource of conserved peptide fragments and residues. *PLoS One* 5, e9679.

García-Jiménez B, Juan D, Ezkurdia I, Andrés-León E, Valencia A (2010). Inference of functional relations in predicted protein networks with a machine learning approach. *PLoS One* 5, e9969.

Leitner F, Krallinger M, Cesareni G, Valencia A (2010). The FEBS Letters SDA corpus: a collection of protein interaction articles with high quality annotations for the BioCreative II.5 online challenge and the text mining community. *FEBS Lett* 584, 4129-4130.

Glaab E, Baudot A, Krasnogor N, Valencia A (2010). Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics* 11, 579.

Tress ML, Valencia A (2010). Predicted residue-residue contacts can help the scoring of 3D models. *Proteins* 78, 1980-1981.

Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A (2010). An Overview of BioCreative II.5. *IEEE/ACM Trans Comput Biol Bioinform* 7, 385-399.

Krallinger M, Leitner F, Valencia A (2010). Analysis of biological processes and diseases using text mining approaches. *Methods Mol Biol* 593, 341-382.

## Awards and Recognition

"Doctor Technices Honoris Causa", Technical University of Denmark

Fellow, The International Society for Computational Biology, USA