Computational Biology

Directed by Prof. Alfonso Valencia, the group is dedicated to the application of machine learning and artificial intelligence to Personalized Medicine, and exhibits ample experience in the development of software platforms for the extraction, integration and representation of big data for large-scale genome projects. The group participates in different international consortia such as the ENCODE project, the ICGC cancer genome effort (including the Spanish subproject on CLL), the BLUEPRINT project and Rare Diseases (RD)-Connect part of the International Rare Diseases Research Consortium (IRDiRC)

Furthermore, Alfonso Valencia is the director of the Spanish National Bioinformatics Institute (INB-ISCIII). The INB represents the Spanish node of the European Bioinformatics infrastructure ELIXIR, and is committed to generating and supplying bioinformatics solutions in the context of national and international activities and consortia.

Objectives

1. Structural Bioinformatics

The study and characterization of structural and dynamical features of protein-protein interactions is of paramount importance in our understanding of cellular mechanisms and the emergence of pathology. Patterns of coordinated mutations that determine changes at protein contacts characterize the co-evolution of interacting proteins. The identification of those mutations allow the systematic prediction of protein contacts and the study of the interplay between co-evolution and structural conservation with a special focus on mutation pathogenicity of specific protein families. Such computational methods devoted to the interpretation of protein variants will expand our knowledge on the mechanisms by which mutations contribute to diseases, including cancer.

- Detection of species-specific coevolution

Co-evolution is fundamental concept for the understanding of evolutionary relationships in ecological networks, protein interactions (Pazos and Valencia, 2001) and even intra-protein contacts. This concept refers to the accumulation of evolutionary changes conditioned by the need to maintain a functional association among evolutionary related entities (species, proteins, etc.). Recent methodological advances in the field of molecular and structural biology allow us to use evidence of strong co-evolution as a reliable marker of important evolutionary relationships at molecular level to better understand and predict fundamental aspects of the structure and function of proteins. With this approach, we have detected differential patterns between somatic mutations previously associated with cancer and polymorphisms detected in the normal population by the 1000 Genomes Project. These results suggest that co-evolution may provide a new vision of the relevance of particular mutations that had previously gone unnoticed.

- Prediction of functional consequences of mutations

Most genomic alterations are tolerated while only a minor fraction disrupts molecular function sufficiently to drive disease. We have developed a solution to facilitate the interpretation of the consequences of mutations for the specific case of human protein kinase variation (a protein family with a central role in cancer), including new methods to predict the pathogenicity of mutations. To understand the biological mechanisms causative of human diseases and cancer, information from pertinent reference knowledge bases and the literature is automatically mined, digested, and homogenized. Variants are visualized in their structural contexts and residues affecting catalytic and drug binding are identified. Known protein–protein interactions are reported.

2. Personalized Medicine

This research line encompasses the development of different strategies and approaches to improved personalized diagnosis of disease, as well as treatment selection for particular patients, based on their individual characteristics. Such a molecular portrait of the patient can be defined, for instance, by integrating genomic information, like mutation profiles and gene expression, as well as other molecular markers. This information can then be used to predict the outcomes of different treatments using information about how other patients with similar characteristics responded to the given treatment. Such enterprise requires a coordinated effort to obtain, process and integrate very heterogeneous sources of information ranging from the sequencing and annotation of genomic variants to the mining of medical literature, as well as the construction of context-specific in-silico models.

- Development of a platform for the management and integration of genomic data from NGS experiments.

The group has developed a comprehensive system for functional analysis of omics data, which has been applied to the data generated by the international projects on cancer genome sequencing ICGC and TCGA, and to datasets obtained by research groups at the CNIO (Epithelial Carcinogenesis Group, directed by Francisco Real; Genetic and Molecular Epidemiology Group, directed by Núria Malats; Translational Bioinformatics Unit, directed by Fatima Al-Shahrour). The analysis platform offers different functionalities such as the prediction of the impact of genomic variation on phenotype, or functional analysis at various levels, including genes, metabolic pathways, genome regions etc. All these features have become accessible through a pioneering web system, which runs analyses and monitors their progress on demand, allowing a level of interaction with the data unprecedented in other similar applications. This line of investigation is related to and has been co-financed by the following projects: ICGC-CLL, BLUEPRINT, RD-CONNECT.

- Analysis of metabolic pathways and regulatory networks associated with the development of various cancers.

Recent genome sequencing studies have shown that somatic mutations that favor the development of cancer spread over a large number of genes. This heterogeneity in the pattern of somatic mutations limits efforts to distinguish sporadic mutations from functional ones. However, based on the widely accepted hypothesis that mutations in cancer affect only a small number of signaling and regulation networks, the study of these networks has become an indispensable tool for analyzing large-scale data ("Cancer Genomics") and discover sets of mutated genes associated with the progression of various tumor types. Studies by our group have highlighted the importance of linking the mutated genes using signaling/regulation and protein-protein interaction networks. We are currently developing a new methodology that allows the systematic analysis of known signaling and regulation networks from data obtained by international consortia (eg. ICGC and TCGA) or recorded in databases of reference (eg. COSMIC). This new methodology is also applied as part of an integrative analysis of the data generated in the International Cancer Genome Consortium (ICGC)'s Pan-Cancer Analysis of Whole Genomes (PCAWG) project.

- Text mining and Natural Language Processing

The overarching benefit of using Text mining to retrieve and analyze biomedical information relies on the possibility of effectively processing large volumes of unstructured knowledge to extrapolate practicable and quantitative content, often in a human-readable fashion. We developed two text mining applications and databases (LiMTox http://limtox.bioinfo.cnio.es/ and MelanomaMine http://melanomamine.bioinfo.cnio.es/), dedicated to the processing of biomedical literature and knowledge resources. In particular, MelanomaMine detects bio-entities (genes, proteins, mutations and chemicals/drugs) that are relevant for the understanding of the molecular basis of diseases. LiMTox extracts associations between compounds and toxicological endpoints facilitating the establishment of toxicity thresholds of several substances. Visibility and precision of our technologies are guaranteed by the platforms OpenMinTteD (http://openminted.eu/), BioCreative (http://www.biocreative.org/), community-wide efforts providing evaluation criteria and guidelines to curate, validate and compare Text mining services, approaches and workflows in terms of interoperability and performance.

- Development of systems biology approaches for modeling and simulating cellular systems.

Computational systems biomedicine relies on the development of in-silico models as a way of integrating different sources of experimental information. Moreover, in-silico simulation of such models produce mechanistic explanations of cellular behavior that can be used, for instance, to design new targeted therapies. In the context of cancer, cell signaling as well as metabolic models have been reconstructed for different cancer types and healthy tissues. In-silico simulation of these models using different computational approaches (e.g. Boolean formalism, Constraint-Based Modeling) have supported the development of targeted therapies that attack specific biological pathways in the cell. These simulations can predict the effect of drugs on cell lines and are being used to understand the process of tumor formation. The approach has mainly been applied in the ASSET and COLOSYS projects. A next step is to integrate different classes of in-silico models into an agent-based model. The aim of developing such a hybrid-model is to study cell dynamics at the population level to understand how cell to cell variability can affect the response to perturbations, such as drug treatments.

3. Big data techniques in transcriptomics s and epigenomics

The group has taken a prominent role in the analysis, integration and interpretation of trascriptomics and epigenomics datasets, especially in the context of the BLUEPRINT project and of the Spanish CLL ICGC project. Understanding how different cell types can be produced by the same DNA sequence requires the disentanglement of various levels of biological regulation. Transcriptomics data is being used to study relationships between patients and cell types at the molecular gene expression level, while a recent focus of the group has been the study of DNA and histone modifications in the 3D nuclear context.

- Disease molecular characterization

Patterns of disease co-occurrence (direct and inverse comorbidity) are being investigated using thousands of transcriptomic datasets to highlight the molecular basis of comorbidity, as part of the EPIC project (MINECO funding). Within the context of the Spanish ICGC CLL project, the group has performed analyses of the transcriptome and epigenome of various CLL patients and related it to healthy B cell differentiation. As part of the BLUEPRINT project, the group has investigated the role of variability in gene expression and DNA methylation in healthy blood cells and CLL patients.

- Epigenomics

The group has played a major role in the BLUEPRINT project devoted to characterising the reference epigenome of healthy haematopoietic cell types. New computational methods were developed for the integration of different epigenomic datasets, including histone modifications, DNA methylation and 3D genome architecture. For example, co-occurrence of different epigenomic factors along the genome was used to identify a network of epigenetic communication. Dimensionality reduction techniques were employed to define chromatin states across the genome in different haematopoietic cell types and to identify the most important genomic regions. Finally a network framework is being used to integrate different kinds of epigenomic datasets onto the experimentally determined 3D structure of the genome.

- Cognitive computing and Artificial Intelligence

A new line of research within the Group is the development of machine learning and cognitive computing solutions to aggregate and analyse complex phenotypic and genotypic data in order to assist personalized medical practices. This project is carried out in collaboration with IBM Academic Initiative (https://developer.ibm.com/academic/) in the framework of the existing agreement between BSC and IBM (BBVA Foundation funding). The overall goal of the project is to predict disease consequences and potential therapeutic interventions using both advanced machine learning techniques, such as Deep Learning, and cognitive computing systems, namely IBM Watson (https://www.ibm.com/watson/) , a technology for reasoning and decision making that combines capabilities in natural language processing, dynamic learning, and hypothesis generation and evaluation.

Source URL (retrieved on *23 Nov 2024 - 10:26*): https://www.bsc.es/es/discover-bsc/organisation/scientific-structure/computational-biology