Eur. Phys. J. Plus (2014) **129**: 130
DOI 10.1140/epjp/i2014-14130-3

THE EUROPEAN PHYSICAL JOURNAL PLUS

Editorial

Pattern recognition tools for proteomics*

Virginio Cantoni[a]

University of Pavia, Dept. of Electrical, Computer and Biomedical Engineering, Via A. Ferrata 5, 27100 Pavia, Italy

Received: 13 February 2014
Published online: 24 June 2014 –

The computer science community started research activities into pattern recognition and artificial vision in the sixties. After so many years of studies and so much research, successfully developed in several fields in recent decades, the pattern recognition community has started to apply the know-how, computing strategies, technologies, methods and tools acquired to new areas, such as computational and structural biology and, in particular, in proteomics.

Some of the lines and perspectives of this initiative are presented in this *Focus Point*. The goal is not a complete survey of the strategies pursued. Rather, it is to describe some remarkable examples of the novel and promising approaches currently under development to the community of physicists.

The selected contributions are related to:

i) *The identification of motifs and domains conserved in families of proteins on the geometric-topologic basis.*

Protein structure analysis and comparison are important to understand the evolutionary relationships among proteins, predicting protein folding and protein functions. A structural motif is a compact 3D protein block, which appears in a variety of molecules. Several motifs are packed together to form domains. Some investigations on protein analysis at various structural levels within a protein or within the entire PDB are discussed and a survey of various approaches to 3D geometrical and topological structure retrieval and comparisons, based on very effective pattern recognition techniques —the Generalized Hough Transform— are presented in detail in the paper *Motifs and structural blocks retrieval by GHT* by Virginio Cantoni, Alessio Ferone, Alfredo Petrosino and Ozlem Polat.

ii) *The prediction of interactions among proteins and other small molecules.*

The identification of protein-binding sites, their classification and analysis are of interest for drug design and treatment of diseases. Binding sites recognition is generally based on geometry and combined with physico-chemical properties, since the conformation, size and chemical composition of the protein surface are all relevant for the interaction with a specific ligand. The amount of work done in this area is huge. In the paper *Predicting protein-ligand and protein-peptide interfaces* by Paola Bertolazzi, Concettina Guerra and Giampaolo Liuzzi a taxonomy of the different approaches is given and their advantages and disadvantages are compared. Broadly speaking, four main categories are envisaged: i) shape-based methods; ii) alignment-based methods; iii) graph-theoretic approaches; and iv) machine-learning methods. In detail, the case of protein-peptide interfaces is considered, in which the binding region peculiarities specialize both geometric and machine-learning methods.

iii) *The volumetric data structure for protein representation and morphological analysis.*

The 3D matching problem is an important objective in the discovery of protein "active sites", *i.e.* complementary regions compatible biochemically, geometrically and topologically, so that they have matching concave and convex segments. The problem is usually pursued by *ad hoc* pattern descriptors which are often point-based and cumbersome for management and processing. In the paper *Structural representation data structures* by Virginio Cantoni, Luca Lombardi, Alessandro Gaggia and Riccardo Gatti, two new main approaches based on different representations (and consequently different data structures) are discussed: the former is based on a hierarchical data structure, the latter on first-order statistics. The first step is the segmentation, performed by classical operators of mathematical morphology, of the protein "solvent-excluded surface" in concavity and convexity regions. Then the interface areas, which can potentially be active sites, are effectively represented through a suitable rich "concavity tree"(CT) and/or through

* Contribution to the Focus Point on "Pattern Recognition Tools for Proteomics" edited by V. Cantoni.

[a] e-mail: virginio.cantoni@unipv.it

an appropriate "extended Gaussian image" (EGI), which represents the histogram of the surface orientations. When we have a large molecule (receptor) and a small molecule (ligand) docking, which take place in a protein cavity, then EGI is certainly suitable. The protein-protein case is usually different; in fact the docking site is larger with planar characteristics rather than a cavity, and the CT approach is usually preferable.

iv) *The integration of protein-protein interaction (PPI) networks with gene expression and 3D structural data to gather temporal and spatial information.*

The study of the evolution of a biological "system" profits substantially from the study of the evolution of "networks" accomplished by combining different sources of genomic and structural domain information to determine novel biological interactions or evolutionary mechanisms. In the paper *On the integration of protein-protein interaction networks with gene expression and 3D structural data: What can be gained?* by Paola Bertolazzi, Mary Ellen Bock, Concettina Guerra, Paola Paci and Daniele Santoni, different points of view on the issue of dynamics in PPI networks are illustrated. A schema is presented concentrating the study of the dynamics either on single proteins or on subnetworks, varying in size and topology, of PPI networks. Finally, some results are given on the dynamics of a PPI network, in which a number of local and also global properties are substantially modified.

v) *The databases annotation with functional information extracted by ontologies and their use to improve the algorithms for the alignment of interaction networks.*

The annotation of biological databases aims to define methods to represent protein interaction data by functional information derived by ontologies (Gene Ontology) and to develop novel algorithms for the comparative analysis of networks in addition to morphological and topological aspects. The functional information may be related to protein localization within the cell, molecular function or the related process. This provides a richer semantics of interaction data and allows concept-based queries, that is, queries based on protein identifiers and, additionally, on their biological characteristics. Moreover, alignment algorithms benefit, too, from the information stored in the protein-protein interaction database for the discovery of unknown interaction and protein complexes or functionalities of not yet completely known proteins. The paper *Annotation and retrieval in protein interaction data bases* by Mario Cannataro, Pietro Hiram Guzzi and Pierangelo Veltri presents two novel software tools, OntoPIN and CytoSeVis, both based on the use of Gene Ontology annotations for the advanced querying of protein interaction databases and for the enhanced visualization of protein interaction networks.

In conclusion, thanks to the efforts of the authors, the selected five papers presented in this *Focus Point*, survey some of the most promising contributions and tools of the pattern community to proteomics.