IL NUOVO CIMENTO Vol. 35 C, N. 5 Suppl. 1 Settembre-Ottobre 2012
DOI 10.1393/ncc/i2012-11329-y

Colloquia: PR PS BB 2011

Preface

V. Cantoni[1] and G. Maino[2]

[1] *University of Pavia - Pavia, Italy*
[2] *ENEA and University of Bologna - Bologna, Italy*

Pattern recognition, after many years of studies and researches successfully developed in several applicative areas, has now know-how, computing strategies, technologies, methods and tools to exploit in new fields such as proteomics, structural biology and bioinformatics.

The amount and complexity of bioinformatics data such as DNA and protein sequences, genetic information, biomedical text and molecular data had a sort of explosion in the past decade. As of Tuesday February 21, 2012 at 4 PM PST there are 79521 experimentally determined 3D structures of protein deposited in the Protein Data Bank (`http://www.rcsb.org/pdb/home/home.do`) with an increment of about 700 new molecules per month.

The importance of studying such huge amount of data, for the analysis of structural building blocks, their comparison and their classification are instrumental to practical problems of the maximum impact, such as the design of a small molecule to bind a known protein or the scan of drugs libraries to detect a suitable inhibitor for a target molecule.

Advanced pattern recognition methods can also play a significant role in high-throughput functional genomics and system biology, where the classification of complex large scale expression profiles, and their link with motif discovery and inference of gene regulatory network, is a major research challenge in the field of Computational Biology.

However, current pattern recognition techniques to tackle these huge data are still not sufficient: The development of approaches for the improvement of the current performances has been the scope of a workshop held in Ravenna, Italy, on September 13, 2011 titled *Pattern Recognition, Proteomics, Structural Biology and Bioinformatics (PR PS BB).* This workshop integrates and continues the tradition of the International Conferences on Image Analysis and Processing (ICIAP 2011, September 14-16), one of the longest running international conferences that started in Italy, in 1980.

The papers here presented, selected from the PR PS BB workshop, can be grouped in five classes, as discussed in the following: i) around the folding problem; ii) protein structural analysis and retrieval; iii) around the docking problem; iv) computational and comparative genomics; v) cells evolution analysis.

Around the folding problem

Several physical-chemical models and computational tools have been developed targeting the folding problem: Having a new sequence of amino acids and a library of known three-dimensional (3D) structures, predict how it folds up in the space. Among the different methods, introduced there are those based on neural networks (NNs), hidden Markov models (HMMs), support vector machines (SVMs), hidden neural networks (HNNs), extreme learning machines (ELMs) and conditional random fields (CRFs), etc. After a survey of these approaches, R. Casadio introduced grammatical-restrained hidden conditional random fields (GRHCRFs). The main GRHCRF novelty is the possibility of including in HCRFs prior knowledge of the problem by means of a defined grammar. In their contribution, R. Casadio, P. L. Martelli, C. Savojardo and P. Fariselli revise a major important problem in bioinformatics: How to annotate protein sequences in the genomic era and all the solutions that have been described by implementing tools based on labelling methods.

The paper of G. Maino shows how it is possible to reduce very complex problems such as those of protein folding or polymer dynamics to a series of manageable steps, by exploiting dynamical symmetries of the considered system and starting from the description of fundamental structural patterns to their mutual interaction and the computation of the dynamics up to the formation of the final equilibrated 3D structure. As a whole, a dramatic reduction in the numerical complexity of the protein folding problem can be attained and in some cases simple solutions can be carried out that provide a convenient physical framework for the understanding of mechanisms responsible for the final 3D structure and the relevant functional properties.

Protein structural analysis and retrieval

Some solutions of the same protein structural comparison problem following disparate inspirations, so leading to diverse approaches, are presented. It turns out that the existence of multiple heuristics for protein structure comparison is actually vital and the reason for this is that it is very difficult to establish quantitatively the distance from optimality of suboptimal structural comparison results. Therefore, only when these methods yield results which are in accordance with one another, despite the fact that the nature of the calculations can be profoundly different, can the predictions be deemed accurate.

In the paper of V. Cantoni, E. Mattia, the Generalized Hough Transform applied to motifs, domains and entire proteins retrieval into a protein data base is introduced. The spatial attitude of a single protein secondary structure (SS) constitutes the item supporting the contributions. For each SS, the locus of contributions (mapping rule) is a circle. There can be spare contributions from neighbors SS, there can be some flexibilities in the block, there can be similarities and not just equalities, there can be finally noise and approximation in the evaluation of the axis of the SS. For all these reasons, it is necessary an analysis of the neighborhoods around the areas with high contributions density (circles intersections). Therefore, both a convenient data structure for effective operations in the neighborhoods (a range tree data structure) and suitable decision criteria have been introduced.

In the paper of V. Cantoni, A. Ferone, A. Petrosino, instead, the spatial distribution of rigid arrangement of protein secondary structures (SSs) constitutes the items supporting the contributions. Starting from the co-occurrence of two not necessarily homogeneous SSs, the approach can be generalized easily up to an entire motif composed of a few SSs. The main characteristic of this approach is that even for the simple couple of SSs, the mapping rule is reduced to a single location for each item of co-occurrence. This reduces very much the signal-to-noise ratio on the parameter space and simplifies the "concrete" data structure, obviously with the drawback of a more elaborated pre-analysis.

In the next paper of V. Cantoni, A. Ferone, R. Oliva, A. Petrosino, the histogram of spatial orientation of rigid arrangement of protein SSs constitutes the item supporting the contributions. A well-known shape representation—usually applied for 3D object recognition—is the Extended Gaussian Image (EGI), which maps on the unitary sphere the histogram of the orientations of the object surface. The adoption of a similar "abstract" data-structure, named Protein Gaussian Image (PGI), for representing the orientation of the protein SSs is proposed. This representation is very effective for a preliminary screening in looking in a protein data base for retrieval of a given structural block, or a domain, or even an entire protein.

Around the docking problem

Molecular docking programs play a crucial role in drug design and development. In the work of P. Bertolazzi, C. Guerra, F. Lampariello and G. Liuzzi is proposed a new docking algorithm which is based on the use of a filling function method for continuous constrained global optimization. Indeed, the protein-peptide correct docking position is sought by minimizing the conformational potential energy subject to constraints that are needed to preserve the primary sequence of the given peptide. The proposed method is based on the idea of modifying the original objective function once a local minimum has been attained, by adding a filling term to it. In this paper it is shown that this algorithm can be profitably used to solve relevant problems in drug design such as the comparison and recognition of protein binding sites and the protein-peptide docking.

The study of the set of protein interactions in a single organism is important for the comprehension of molecular processes. The possibility to annotate such data using Gene Ontology and the use of semantic similarity measures facilitate and enhance the developing of novel algorithms for protein interactions analysis. However, semantic similarity measures may be affected from some biases. The paper of M. Mina and P. H. Guzzi demonstrates the existence of the bias that affects main semantic similarity on a set of well-known yeast complexes. It also provides some evidence about the variability of the bias effects over the proteome.

Protein-protein interaction takes usually place on extended areas of the molecules' surfaces that are morphologically fitting. It is therefore important to adopt representations and data structures that can make the analysis easier as well as the implementation of techniques for the evaluation of geometric and topological properties on extended surfaces. These areas of activity are usually roughly "planar" but with local concavity and complexity that must match for interacting. The solution suggested by V. Cantoni, R. Gatti and L. Lombardi is based on the concavity tree representation. Each node of the tree contains a vector of features that describes the geometrical, topological and biochemical properties of the corresponding surface patch.

Referring to protein ligand interaction, the "active sites" are always located in one of the biggest concavities (in one of the larger 'pockets') and the ligand must match this concavity, so its effective part must be mainly convex. For this reason, in the paper of V. Cantoni, A. Gaggia and L. Lombardi, the matching potential can be evaluated through an Extended Gaussian Image (EGI) shape representation. The original EGI, and a few extensions (namely, Complex EGI and Enriched Complex EGI) representations and their correspondent concrete data-structures are discussed and exploited for the evaluation of the matching aptitude between the small ligand molecule and a pocket of a protein macromolecule.

Computational and comparative genomics

The changes' monitoring at DNA level enables the characterization of the underlying structure of genetic diseases. Hence, the development of algorithms aimed at the identification of copy number alterations (CNAs) is a current challenge in Bioinformatics. Despite the amount of proposed approaches, identification of CNAs is yet an open problem. The work of S. Morganella and M. Ceccarelli proposes a new algorithm that starts from the assumption that copy number profiles are piecewise constant and finds the optimal segmentation by minimizing a functional energy that represents a compromise between accuracy and parsimony of the boundaries.

A hot topic of actual research in molecular biology is the study of microRNA-gene interactions. The paper of S. Rovetta, F. Masulli and G. Russo evaluates the performances a data-oriented method by means of a cross-validation approach. Latent information can be effectively exploited in order to suggest directions for laboratory experiments, an important topic in microRNA research, since these experiments are costly in both resources and time.

P. F. Stifanelli, T. M. Creanza, R. Anglani, V. C. Liuzzi, S. Mukherjee and N. Ancona present a comparative study of Gaussian Graphical Model (GGM) approaches for genomic data. In particular, in this paper the focus is on regularized methods for the estimation of the concentration matrix in an undirected GGM. A comparative study of three methods to obtain small sample and high dimension estimates of partial correlation coefficients has been performed. Even if two regularized methods show comparable best performances, the experimentation shows that the covariance-regularized method has to be preferred because of the shorter computation time.

Cells evolution analysis

Gene circuit dynamics can be precisely estimated at a single-cell level through automatic cell segmentation and tracking. This procedure is particularly important because of the large number of cells usually under study and the huge amount of images to be analyzed. In the paper of R. La Brocca, F. Menolascina, D. di Bernardo and C. Sansone, the analysis of yeast cells in bright field and phase-contrast microscopy images is presented. The solution is shown to be robust to experimental variability and able to automatically find the best set of parameters via a pattern recognition approach and therefore it can be used by biologists with little knowledge in the field of image processing, even achieving competitive results when compared with best case scenarios of alternative solutions.