Computational Intelligence and Machine Learning Methods in Bioinformatics

Giorgio Valentini [a] Roberto Tagliaferri [b] Francesco Masulli [c,d]

[a] *DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano, Via Comelico 39, Milan, Italy.*

[b] *DMI, Dipartimento di Matematica e Informatica,
Università di Salerno, Via Ponte don Melillo, Fisciano (SA), Italy.*

[c] *DISI, Dipartimento di Informatica e Scienze dell' Informazione,
Università di Genova, Via Dodecaneso 35, Genoa, Italy.*

[d] *Center for Biotechnology, Temple University, 1900 N 12th Street Philadelphia,
PA 19122, USA.*

1 Introduction

The treatment of huge amounts of data delivered by high-throughput biotechnologies requires on one hand advanced data management procedures for an efficient storage and retrieval of biological information [1], and on the other hand refined methods to extract and model biological knowledge from the data [2].

Computational intelligence (CI) and machine learning (ML) methods are widely applied to the extraction of biological knowledge from bio-molecular data [3,4], in order to obtain models to both represent biological knowledge and to predict the characteristics of biological systems.

Email addresses: `valentini@dsi.unimi.it` (Giorgio Valentini), `rtagliaferri@unisa.it` (Roberto Tagliaferri), `masulli@disi.unige.it` (Francesco Masulli).

It is worth noting that a well-known machine learning algorithm (namely, the perceptron, inspired to the behaviour of a neuronal cell [5]), has just been applied to bioinformatics in the eighties to distinguish translation initiation sites in prokaryotic organisms [6], and starting from this early application, a growing number of computational intelligence and machine learning methods has been applied and often developed to deal with a wide range of bioinformatics problems in genomics, proteomics, gene expression analysis, biological evolution, systems biology, and other relevant bioinformatics domains.

Genomics studies biological sequences at genome level. CI and ML methods have been applied to the reconstruction and sequencing of entire genomes [7,8], to the extraction and identification of the structure of genes [9,10], to the identification and analysis of regulatory non coding DNA elements [11,12], to the genome-wide identification of genes involved in genetic diseases [13], to the prediction of phenotypic effects of non synonymous single nucleotide polymorphisms [14], to identify RNA structural elements [15], to model haplotype blocks [16], to splice site prediction [17], to the detection of gene to gene interactions in studies of human diseases [18], to multiple alignment of bio-sequences in phylogenomics [31], and to many other relevant genomics problems.

In proteomics the main problem of the prediction of the secondary and tertiary structure of proteins represent one of the main challenges for CI and ML methods in bioinformatics [19]. Another key problem in proteomics (and in genomics too) is the prediction of the functions of the proteins and genes: the large scale sequencing programs make available sequences of entire genomes of several organisms, but besides the identification of genes, we need to understand their properties and the functions of the corresponding proteins [20,21]. Many other problems in proteomics have been formalized as machine learning problems, such as fold recognition [22], the prediction of contact maps [23], and protein subcellular localization [24].

Gene expression data analysis and in particular transcriptomics is another well-established bioinformatics domain where CI and ML methods have been widely applied, providing significant results in several fields of molecular biology and medicine [25,26]. Three kinds of problems have been mainly studied within the community of bioinformaticians for answering three basic questions [27]: a) Class prediction, that is the determination of the functional state of a cell or tissue through the expression level of its genes [28,29]; b) Gene selection: the identification of genes correlated to the functional state under investigation [4]; c) Class discovery: analysis of the groups (clusters) of co-expressed and functionally correlated genes/examples [30].

Systems biology is an emerging bioinformatics area [32] where ML and CI techniques play a central role. Indeed modeling biological processes inside cells, and more in general biological systems, require the development of mathematical models and learning methods to fit the models to biological data. In particular, probabilistic graphical models have been widely applied to model biological networks [33], ranging from genetic [34] to metabolic [35] and signal transduction networks [36].

In other relevant bioinformatics and bioinformatics-related areas, such as biomedical image analysis, ML and CI methods have been successfully developed and applied, but of course a thorough overview of this so wide and growing research area is far beyond the scope of this editorial.

2 Special issue contents

The special issue presents 13 papers with contributions coming from diverse areas of machine learning and computational intelligence methods for bioinformatics. The papers have been selected after extensive reviews and revisions, starting from about 50 papers submitted to the Fourth International Meeting on *Computational Intelligence methods for Bioinformatics and Biostatistics* (CIBB 2007) that was held in Portofino Vetta, Ruta di Camogli (Italy) in July 2007 in the framework of the activities of the Special Interest Group in Bioinformatics of the International Neural Network Society. The main goal of CIBB meetings is to provide a forum open to researchers from different disciplines to present and discuss problems relative to computational techniques in bioinformatics and medical informatics with a particular focus on machine learning and computational intelligence methods.

The papers of this special issue embrace a wide range of bioinformatics areas. In a glance, the papers are subdivided in three main groups, according to three of the main general bioinformatics domains: genomics, transcriptomics and proteomics. Even if for some papers this subdivision is quite schematic (indeed several research areas embrace different bioinformatics domains), we follow this broad taxonomy according to the main subdivisions of bioinformatics research areas adopted by machine learning and computational intelligence communities [3,4].

The opening paper is the extended version of the *Joaquin Dopazo* invited talk at CIBB 2007 and provides a general overview of a new research line in functional genomics. Dopazo's paper is included in the genomics section, but considering its wide-ranging domain, it could be included in the transcriptomics section as well.

2.1 Genomics

Joaquin Dopazo [37] provides a thorough critical dissertation on the hypothesis formulation and testing in functional genomics, introducing new perspectives for the development of computational methods that relate the available genomic information with the hypotheses that originated the experiments. In this paper the author reviews the main characteristics of functional enrichment methods, by which we can find if gene modules, that is groups of genes related by some related biological property (e.g. Gene Ontology functional modules [38] or KEGG pathways [39]), are significantly overrepresented among the relevant genes selected in the experiment. The inconsistencies in the way functional hypotheses are tested by functional enrichment methods are analyzed, and new methods, generally known as gene sets analysis methods, inspired by systems biology criteria, are critically discussed and reviewed. These methods recently proposed in the domain of functional genomics, are based on the biological fact that modules, and not single genes constitute the ultimate functional "bricks" which act cooperatively to carry out functions in the cell. In particular new supervised and unsupervised methods that attempt to exploit "a priori" biological knowledge of functional relationships between sets of genes (modules) are discussed, as well as applications of gene sets analysis in transcriptomics, large scale genotyping and phylogenomics.

Michele Ceccarelli and *Alessandro Maratea* [40] address the problem of the alternative splicing prediction, a key mechanism to understand the multiplicity of proteins raising from a relatively low number of genes in eukaryotic organisms. The authors present a supervised machine learning approach using support vector machines with data obtained from a virtual genetic coding scheme to numerically modeling the information content of sequences, and using time series analysis to extract fixed-length set of features from each sequence. Machine Learning recognition of alternatively spliced Exons reaches an AUC of over 96% on tested *C. Elegans* data, confirming to be an effective procedure especially when no "a priori" biological knowledge is available. As a byproduct of this study the virtual genetic code based on Shannon information content proposed in this paper could be an attractive option whenever a numerical translation of a biological sequence is needed, and could be in principle applied in other areas of genomics and transcriptomics.

Matteo Rè and *Giulio Pavesi* [48] address the problem of the detection of the conserved coding genomic regions through signal processing techniques applied to the analysis of the alignment of nucleotide sequences of different organisms at the level of the entire genome. The authors analyze the DFT (Discrete Fourier Transform) spectrum of the signal of the mismatches between two human/mouse aligned sequences of length N. The main idea behind this approach consists in the biological fact that coding regions mismatches occur predominantly in the third codon position, while they should appear almost randomly in non-coding regions, thus resulting in a higher N/3 frequency component in coding regions. The authors propose measures based on this analysis to unravel the coding potential of genomic regions. This method, that applies signal processing methods in a comparative genomics framework, can be in principle extended to the analysis of the genome of other organisms, because of the universality of the genetic code and the selective pressure acting on protein coding regions.

Luca Nicotra and *Alessio Micheli* [47] tackle the problem of gene function prediction using phylogenetic data. To this end they propose supervised learning methods based on a a class of kernels for structured data leveraging on a hierarchical probabilistic modeling of phylogeny among species: a sufficient statistics kernel, a Fisher kernel, and a probability product kernel. The authors introduce kernel adaptivity to the data through the estimation of the parameters of a tree structured model of evolution, showing an improvement in the classification of functional classes of genes in *S. Cerevisiae* w.r.t. to standard vector based kernel and non-adaptive tree kernel functions.

Alessandro Perina, Matteo Cristani, Luciano Xumerle, Vittorio Murino, Pier Franco Pignatti and *Giovanni Malerba* [51] address two central problems in medical genetics, related to the localization of genetic regions containing susceptibility genes for genetic diseases: the haplotype reconstruction and haplotype block discovery. To this end they propose a new Hidden Markov model (HMM) and an inference strategy for learning. The estimation of haplotypes from genetic patterns in unrelated individuals is performed by applying variational learning strategies, thus avoiding local minima solutions that affect other HMM methods based on the classical Expectation-Maximization algorithm. Moreover the proposed Fully Non Homogeneous HMM is able to segment genotypes into linkage disequilibrium blocks, using the Gini index, a classical statistical measure, to determine the segmentation of block boundaries. The results are competitive with state-of-the-art systems for haplotype reconstruction and block discovery.

2.2 Transcriptomics

Oleg Okun and *Helen Priisalu* [41] present a paper that opens a new approach to computer-aided bio-molecular diagnosis of malignancies, explicitly taking into account the complexity that characterizes high-dimensional gene expression and other types of bio-molecular data. The authors propose a supervised ensemble method based on the complexity of the data obtained by randomly choosing subsets of features (gene expression levels associated to specific genes) and then selecting only the least complex data through a proper measure of complexity. The authors show also through an extensive statistical analysis that there is a direct relationship between the accuracy of the base learners (estimated through a low biased bolstered error) and the complexity of data (estimated through an adaptation of the Wilcoxon rank sum test). The proposed new scheme for generating ensembles of classifiers is applied to the analysis of several gene expression data sets, showing that the selection of features/genes leading to less complex data ensures a better performance of the resulting ensemble.

Weightless connectionist models in which each neuron performs basically boolean operations and analog to digital conversions are proposed by *Massimiliano Costacurta, Marco Muselli* and *Francesca Ruffino* [42], in conjunction with Recursive Feature Addition (RFA) techniques to properly select genes related to a specific phenotype. By this technique the authors are able to assign a relevance value to the variables associated to the expression level of each gene and to select the most relevant through the RFA approach. The effectiveness of the method is demonstrated by using a recently proposed mathematical model based on the biological concepts of expression signature and expression profile on both real and artificial gene expression data.

In the paper presented by *Roberto Avogadri* and *Giorgio Valentini* [44], the authors address the problem of the uncertainty underlying the assignments of examples/patients to clusters in the context on unsupervised gene expression data analysis. This problem is relevant to discover subclasses of pathologies based on the bio-molecular characteristics of patients. To deal with this problem, a fuzzy approach is adopted by applying a fuzzy-k-means algorithm to different instances of the data and by using a fuzzy t-norm to combine the multiple clusterings. The multiple instances of the data are obtained by Bernoulli random projections that reduce the high dimensionality of gene expression data, without introducing relevant distortions into the data, thus improving both the accuracy and the diversity of the obtained base clusterings. The advantages and limitations of the proposed approach are shown by comparing its accuracy and robustness w.r.t. state-of-the-art clustering ensemble algorithms. Finally, an empirical analysis of the relationships between the accuracy and diversity of the base fuzzy-clusterings is provided.

The paper of *Paola Campadelli, Elena Casiraghi* and *Andrea Esposito* [50] is included in this section only for organizational reasons, but its research domain is within bio-medical image analysis, an important research area related to bioinformatics, especially in the perspective of the integration of multisource biological data for the diagnosis and outcome prediction of diseases. The paper provides a description and a critical analysis of the state of the art of semi-automatic and automatic liver segmentation techniques and of a new algorithm to deal with most of the problems and drawbacks of the computational methods discussed in the review. Live wire segmentation approaches, gray level based methods, neural networks, Bayesian and model fitting based methods are reviewed, in order to analyze the pros and cons of different image processing methods that constitute the first step for the automatic liver disease diagnosis and three-dimensional liver rendering. The authors propose a three-steps gray level based technique to cope with the high inter and intra patient gray level and shape variability, achieving a high accuracy in the liver segmentation obtained from 40 abdominal contrast enhanced computed tomography images.

2.3 Proteomics

Gennady Verkhivker's paper [43] focuses on the problem of the analysis of binding mechanisms and molecular signatures of the HIV-1 protease drugs. HIV-1 PR represents an important target for the design of antiviral agents, and in this work the molecular basis of the HIV-1 PR inhibition are studied. To this end Monte-Carlo simulations with the conformational ensembles of the HIV-1 PR dimer and monomer structures have been performed, thus enabling a molecular analysis of the active site and dimerization modes of inhibition. The author shows that an acetylated tetrapetide Ac-SYEL-OH can act as both a dimerization inhibitor and a competitive active site inhibitor, and unravels the way that the peptide NIIGRNLLTQI acts as folding inhibitor of HIV-1 PR, thus enabling the design of novel inhibitors of HIV-1 protease.

The classification of protein samples w.r.t. a given phenotype is one of the major goals in quantitative proteomics. When comparing two biological samples measured with liquid chromatography coupled to mass spectrometry (LC/MS), one often observes a nonlinear time deformation between consecutive experiments which introduces a severe alignment problem.

Bernd Fischer, Volker Roth and *Joachim M. Buhmann* [45] address this problem by applying a method based on Generalized Canonical Correlation Analysis, in order to improve the estimation of differential protein expression values. In particular they introduce an adaptive scale space estimation for complexity tuning of the time-warping functions, and a local model selection procedure for each time axis instead of the usual global model selection procedure. With this novel technique the authors overcome two severe problems of the previous approaches: non-symmetry of the time prediction function and a potential violation of the monotonicity constraint in temporal alignments.

The classification of high-dimensional mass-spectrometry measurements represents a challenging CI and ML problem, with significant applications in cancer research.

Frank-Michael Schleif, Thomas Villmann, Markus Kostrzewa, Barbara Hammer and *Alexander Gammerman* [46] propose a supervised prototype based classifier applied to mass spectrometric data preprocessed with wavelets techniques that uses a functional norm that takes into account the specific nature of mass-spectra. The authors propose as prototype based classifier the Supervised Relevance Neural Gas (SRNG) whose accuracy, in this context, is comparable with state-of-the-art supervised learning algorithms. Moreover, considering that SRNG generates models which consist of typical points of the data, prototypes for the classes under investigation, the solution representation allows also a more natural interpretation of data from a bio-medical standpoint.

Marco Vassura, Luciano Margara, Piero Fariselli and *Rita Casadio* [49] address the problem of Protein Structure Selection (PSS), that is the assignment of a given protein to one of 3D structures (named decoys) according to a given distance measure. In literature, existing methods for solving PSS usually rely on primary structure of the protein and on protein chemistry, making use of specific energy functions that need to be minimized through suitable optimization methods. On the contrary, the authors propose an original approach to the selection of decoys which are closer to the original (unknown) structures, based solely on geometric and graph-based information. Indeed, they show that graph properties can be used to assess the quality of a prediction of the native state structure of a protein, identifying important connections between properties of decoys and graphs. The results show that, based on simple geometrical properties, graph-based predictions can be as robust as seemingly more sophisticated energy-based scoring of best decoys, opening new perspectives for solving PSS problems.

Acknowledgements

We would like to thank all the reviewers for their critical, yet constructive, comments that significantly improved the quality of the papers. We would also like to thank Klaus-Peter Adlassnig, who made possible the publication of this special issue.

References

[1] C. Goble, R. Stevens, State of the nation in data integration for bioinformatics., Journal of Biomedical Informatics (in press), available on line at http://www.sciencedirect.com.

[2] P. Baldi, S. Brunak, Bioinformatics, The machine learning approach, MIT Press, Cambridge, MA, 2001.

[3] K. Cios, H. Mamitsuka, T. Nagashina, (Eds.), Computational intelligence techniques in bioinformatics (Special issue), Artificial Intelligence in Medicine 35 (1-2) (2005).

[4] P. Larranaga, B. Clavo, R. Santana, C. Bielza, J. Gladiano, I. Inza, J. Lozano, R. Armananzas, G. Santafe, A. Perez, V. Robles, Machine learning in bioinformatics, Briefings in Bioinformatics 7 (1) (2006) 86–112.

[5] F. Rosenblatt, The perceptron, a probabilistic model for information storage and organization in the brain, Psychological Review 65 (1958) 386–408.

[6] G. Stormo, T. Scheider, L. Gold, A. Ehrenfeuch, Use of the perceptron algorithm to distinguish translation initiation sites in E.coli, Nucleic Acid Research 10 (1986) 2997–3011.

[7] E. Lander, et al., Initial sequencing and analysis of the human genome, Nature 409 (6822) (2001) 860–921.

[8] J. Venter, et al., The sequence of the human genome, Science 291 (5507) (2001) 1304–1351.

[9] M. Brent, R. Guigo, Recent advances in gene structure prediction, Current Opinion in Structural Biology 14 (3) (2004) 264–272.

[10] A. Bernal, K. Crammer, A. Hatzigeorgiou, F. Pereira, Global discriminative learning for higher−accuracy computational gene prediction, PLoS Computational Biology 3 (3) (2007).

[11] G. Ratsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.-R. M??ller, R.-J. Sommer, B. Scholkopf, Improving the caenorhabditis elegans genome annotation using machine learning, PLoS Computational Biology 3 (2) (2007) e20.

[12] D. Holloway, M. Kon, C. DeLisi, Machine learning for regulatory analysis and transcription factor target prediction in yeast, Systems and Synthetic Biology 1 (1) (2007) 25–46.

[13] N. Lopez-Bigas, C. Ouzounis, Genome-wide identification of genes likely to be involved in human genetic diseases, Nucleic Acid Research 32 (10) (2004) 3108–3114.

[14] L. Bao, Y. Cui, Prediction of the phenotypic effects of non synonymous single nucleotide polymorphisms using structural and evolutionary information, Bioinformatics 21 (5) (2005) 2185–2190.

[15] G. Fogel, W. Porto, D. Weekes, Prediction of the phenotypic effects of non synonymous single nucleotide polymorphisms using structural and evolutionary information, Nucleic Acid Research 30 (23) (2002) 5310–5317.

[16] G. Greenspan, D. Geiger, High density linkage disequilibrium mapping using models of haplotype block variations, Bioinformatics 20 (S1) (2004) 137–144.

[17] Y. Saeys, et al., Feature subset selection for splice site prediction: a new method using eda-based feature ranking, BMC Bioinformatics 5 (64).

[18] M. Ritchie, et al., Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases, BMC Bioinformatics 4 (28).

[19] A. Randall, J. Cheng, M. Sweredoski, P. Baldi, Tmbpro: secondary structure, beta-contact and tertiary structure prediction of transmembrane beta-barrel proteins, Bioinformatics 24 (4) (2008) 513–520.

[20] O. Troyanskaya, et al., A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *saccharomices cerevisiae*), Proc. Natl Acad. Sci. USA 100 (2003) 8348–8353.

[21] Z. Barutcuoglu, R. Schapire, O. Troyanskaya, Hierarchical multi-label prediction of gene function, Bioinformatics 22 (7) (2006) 830–836.

[22] A. Raval, Z. Ghahramani, D. Wild, A Bayesian network model for protein fold and remote homologue recognition, Bioinformatics 18 (6) (2002) 788–801.

[23] G. Pollastri, P. Baldi, Prediction of contact maps by giohmms and recurrent neural networks using lateral propagation from all four cardinal corners, Bioinformatics 18 (S1) (2002) 62–70.

[24] Y. Huang, Y. Li, Prediction of protein subcellular localization using fuzzy k-nn method, Bioinformatics 20 (1) (2004) 21–28.

[25] D. Allison, X. Cui, G. Page, M. Sabripour, Microarray data analysis: from disarray to consolidation and consensus., Nat Rev Genet. 7 (1) (2006) 55–65.

[26] S. Wang, Q. Cheng, Microarray analysis in drug discovery and clinical applications., Methods Mol Biol. 316 (2006) 49–65.

[27] J. Dopazo, Functional interpretation of microarray experiments, OMICS 3 (10) (2006).

[28] S. Dudoit, J. Fridlyand, T. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, JASA 97 (457) (2002) 77–87.

[29] Z. Lee, An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer, Artificial Intelligence in Medicine 42 (1) (2008) 81–93.

[30] J. Handl, J. Knowles, D. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics 21 (15) (2005) 3201–3215.

[31] J. Handl, D. Kell, J. Knowles, Multiobjective optimization in bioinformatics and computational biology, IEEE/ACM Trans. Comput. Biol. Bioinformatics 4 (2) (2007) 279–292.

[32] H. Kitano, Systems biology: A brief overview, Science 295 (5560) (2002) 1662–4.

[33] J. Bower, H. Bolouri, Computational Modeling of Genetic and Biochemical Networks, MIT Press, 2004.

[34] N. Friedman, Inferring cellular networks using probabilistic graphical models, Science 303 (2004) 799–805.

[35] M. Green, P. Karp, A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases, BMC Bioinformatics 5 (76) (2004).

[36] M. Steffen, et al., Automated modelling of signal transduction networks, BMC Bioinformatics 3 (34) (2002).

[37] J. Dopazo, Formulating and testing hypothesis in functional genomics, Artificial Intelligence in Medicine (in this issue).

[38] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology, Nature Genet. 25 (2000) 25–29.

[39] M. Kanehisa, S. Goto, Kegg: Kyoto encyclopedia of genes and genomes, Nucleic Acid Res. 28 (2000) 27–30.

[40] M. Ceccarelli, A. Maratea, Virtual genetic coding and time series analysis for alternative splicing prediction in C.Elegans, Artificial Intelligence in Medicine (in this issue).

[41] O. Okun, H. Priisalu, Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors, Artificial Intelligence in Medicine (in this issue).

[42] M. Costacurta, M. Muselli, F. Ruffino, Evaluating gene selection methods through artificial and real gene expression data, Artificial Intelligence in Medicine (in this issue).

[43] G. Verkhivker, Computational proteomics analysis of binding mechanisms and molecular signatures of the HIV-1 protease drugs, Artificial Intelligence in Medicine (in this issue).

[44] R. Avogadri, G. Valentini, Fuzzy ensemble clustering based on random projections for DNA microarray data analysis, Artificial Intelligence in Medicine (in this issue).

[45] B. Fischer, V. Roth, J. Buhmann, Adaptive bandwidth selection for biomarker discovery in mass spectrometry, Artificial Intelligence in Medicine (in this issue).

[46] F. Schleif, T. Villmann, M. Kostrzewa, B. Hammer, A. Gammerman, Cancer informatics by prototype networks in mass spectrometry, Artificial Intelligence in Medicine (in this issue).

[47] L. Nicotra, A. Micheli, Modeling adaptive kernels from probabilistic phylogenic trees, Artificial Intelligence in Medicine (in this issue).

[48] M. Re, G. Pavesi, Detecting conserved coding genomic regions through signal processing of nucleotide substitution patterns, Artificial Intelligence in Medicine (in this issue).

[49] M. Vassura, L. Margare, P. Fariselli, R. Casadio, A graph theoretic approach to protein structure selection, Artificial Intelligence in Medicine (in this issue).

[50] P. Campadelli, E. Casiraghi, A. Esposito, Liver segmentation from CT scans: a survey and a new algorithm, Artificial Intelligence in Medicine (in this issue).

[51] A. Perina, M. Cristani, L. Xumerle, V. Murino, P. Pignatti, G. Malerba, FNH-HMM double net: a Bayesian network for haplotype reconstruction and haplotype block discovery, Artificial Intelligence in Medicine (in this issue).