



Bioinformatics and Pattern Recognition Come Together

“ - Do you know everything, Jeeves?
 - I really don't know, sir.”

P. G. Wodehouse

After the completion of the sequencing of the human genome, the scientific community made the announcement that we had entered a period of what is now commonly known as a post-genomic era [1,2]. This essentially means that the availability of the drafts of the genome of several organisms (and most importantly of the human genome), has enabled us to study systematically the relationships between genes encoding a programme of life and physiological processes within organisms. Such studies may involve several types of information of varied complexity, which describe a wide range of biological processes or objects. For instance, these can be data about DNA sequences, genetic polymorphisms, protein structures, gene expression levels or protein-protein and protein-DNA interactions, etc. The field of biology is still much at the stage of intensive data gathering and the main analytical breakthroughs are apparently yet to come. A global effort of biologists studying simultaneously tens of thousands of interdependent entities on a molecular level, by means of newly available high-throughput experiment platforms, has created an explosive growth of the amount of raw data. Analysis of these data, allowing for accurate modelling of biological processes at the levels of molecules, cells, tissues and organisms, requires efficient systems for data collection, for standardization and for representation to academic as well as to commercial users. The solution has come with the development of large curated databases. Their fusion with the data analysis has resulted in the creation of *Bioinformatics*.

As our primary focus will be on applications of pattern recognition in various knowledge domains, and in this section - applications in Bioinformatics - we shall briefly guide our readers and potential authors through the terminology of the two fields in order to avoid major confusions and misinterpretations and to present our vision and the scope of the section. Pattern recognition, one of the fields of *machine learning*, is defined by the Encyclopaedia Britannica as “the imposition of identity on input data, such as speech, images, or a stream of text, by the *recognition* and delineation of *patterns* it contains and their relationship”. Methods of pattern recognition are used to classify recurring patterns in data on the basis of either *a priori* knowledge or information directly extracted from the data. Thus, the resulting learning strategy is characterised as either *supervised* or *unsupervised* learning.

Normally, pattern recognition involves gathering observations on certain objects in a standardised manner and classifying them on the basis of a formerly agreed set of object features. Supervised and unsupervised learning are different in the sense that in the first approach relationships between features and classes are derived from the patterns that have already been classified, whereas in the second approach the classes are determined on the basis of statistical information extracted directly from the analysed data. That is why, in bioinformatics, these approaches are often linked to *class prediction* and *class discovery* respectively.

Bioinformatics is an extremely multidisciplinary field which integrates the methods and approaches of molecular biology, computer sciences, physics and physical chemistry and

mathematical modelling. The major current task of the field is still to collect and store biological data. However, together with databases of primary information (such as sequence or structure information) there have emerged resources which provide the results of various meta-analyses. These, for instance, are the databases of protein families and domains, protein structural classes or signalling pathways. Vast amounts of information to be analysed have initiated the development of automated data processing pipelines allowing for primary resources to be accompanied by structured collections of core biological knowledge that may be inferred, from the stored data, by means of various automated or semi-automated analytical methods. The integrative parts of such pipelines are the methods of either supervised or unsupervised classification. It is not rare now to see the results of classification which have been provided by one source become input data for the resources providing analysis on a higher level, e.g. the InterPro integrated resource and its member databases [3]. There are also attempts to unify data related to a particular disease, classified not only according to the technology used for producing the data, but also according to clinical information [4]. On the one hand, obtaining information about biological systems at the level of the transcriptome, proteome or metabolome is vital for progress in the field; on the other hand, it calls for novel analytical methods that could be simultaneously applied to a variety of data types. Thus, data warehousing and retrieval as well as effective communication of the experimental results are areas in dire need of the extensive application of efficient machine learning techniques.

Recently, in the field of bioinformatics there has been an increasing movement to develop new methods for the analysis of collected data, with the ultimate purpose of creating relevant models of biological processes and understanding their underlying mechanisms. We are still very far from the stage where, even simple, core models are developed and widely used. This task is complicated by the very fact that the numbers of objects participating in studied processes usually range here from tens to thousands to tens of thousands, e.g. the number of genes in the genome of the yeast *S. cerevisiae* is about 6000. The data on such systems are of high dimensionality and are often imbalanced and inherently noisy. Thus, gargantuan part of current research is still focused on the formulation of relevant hypotheses regarding studied objects, in order to follow them up with new experiments and analyses. The best example here is probably the study of gene expression or protein interaction data, aiming at elucidating unknown members of cell signalling pathways and gradually uncovering the interplay of participants in biochemical networks in living cells. The development of high-throughput technologies, and microarrays especially, has shaped the aims and purposes of the life sciences. Whereas molecular biology has traditionally worked with selected genes or proteins, sequentially identifying their functions, microarrays have made possible simultaneous screening of whole genomes. This methodology allows quantification of an entire transcriptome (set of all mRNAs that are transcribed in a given cell type) and, recently, a proteome (set of all proteins that are translated from mRNAs) under particular conditions. This has made possible rapid global comparisons, accelerating proliferation of bio-discoveries into clinical studies and ultimately into new drugs and diagnostics. Microarrays help to monitor global gene expression by measuring mRNA abundance levels, providing snapshots of expression levels for the whole system under a particular condition. Further, these data are analysed on the basis of a hypothesis that the dynamics of mRNA abundances of different genes reflect signalling and regulation interrelations between members of the global biochemical network. The common assumption here is that the whole regulatory system may be divided into rather small parts (functional modules), connected to each other via the global signalling network [5-7]. It is quite obvious that only

a small fraction of all “theoretically possible” states of such a network may be attained by a real biological system [8]. Living cells, in practice, cannot be forced to exist under all possible conditions for the simple reason that they may stop functioning, i.e. die at that point. It is very likely that many states are never seen in our experiments. Because of this, we cannot learn the network’s precise interconnections by observing all conceivable states. Besides, the number of global expression snapshots is often significantly smaller than the number of genes or proteins, usually tens or hundreds vs. tens of thousands. Moreover, expression data are noisy by its very nature, because of innate variations of living systems which constantly adapt to environment changes or just go through their normal life cycles. Thus, the best strategy is to construct some workable hypotheses for further experimental verification. To derive such hypotheses from imprecise data and to ensure their quality is exactly the task of machine learning. Thus, a range of induction, feature selection and validation algorithms have recently found their applications in microarray informatics.

Closely related to the above mentioned research is the fast-growing area of machine learning, both supervised and unsupervised, which is used for discrimination between different sample classes or experimental conditions on the basis of gene/protein expression and metabolic profiles. This area includes recognition of cancer classes, verification of disease diagnosis, prediction of treatment and survival outcomes, etc. [9-11]. Pioneering works in cancer classification have laid the ground for entire new research branches in biomedical informatics [9]. Here the methods of class discovery and class prediction have become real eye-openers, revealing that some diseases, hitherto homogeneously characterised, are subdivided into different classes that should actually be approached with different treatment tactics. Another emerging area for the use of supervised methodologies, where unsupervised methods have already been extensively applied, is the learning of functional classification of uncharacterised genes or proteins [12-14]. One of the common approaches here is to identify sequence motifs, combinations of which accurately predict functional and some of the structural properties of uncharacterised sequences. A variety of sequence analysis methods has initiated the development of methods for integrated annotation. Sequence motifs are inferred by using various statistical models and techniques, e.g. HMMs, clustering, etc. [3]. The need to account for all relevant sources and to produce cross-method consistent classifications is the focus of the machine-learning experts specializing in automated integration of the biological information supplied by various sources [12,15-16]. Feature selection algorithms are designed for the situations where large volumes of diverse data are to be analysed together, a task that is usually impossible to undertake manually. Selection of relevant features needs to be linked to resulting functional characteristics, which poses other complex problems. In particular, the task of translating sequence-related data into functional annotation is augmented by the existence of a wide range of function descriptors (e.g. GO terms [17], SwissProt keywords [18], etc.). Given that neither of the latter is absolutely accurate nor complete, the challenge is to complement expert-driven curation of protein data by apt pattern recognition algorithms.

I think I have now reached the point where it has become clear what kind of studies we are seeking to have in this section of the journal. If I were to answer the question posed to P.G. Wodehouse’s character in the preface to this piece, my answer would have been a definitive “No, I do not!” Thus, taking into account that the length of this foreword should be rather short, I shall wind up. Otherwise, I would start to disseminate information available elsewhere, while pretending that I am saying something new. Thus, all those who are interested in the state of the art of a particular area in bioinformatics I shall redirect to available specialised reviews. My only remaining concern is that I did not cover all or even

the majority of areas in the fascinating field of bioinformatics, where applications of pattern recognition are relevant. Therefore, I would like, wholeheartedly, to invite those who work in genomics, proteomics, metabolo- and metabonomics and indeed all other -omicses; those who try to systematize biological knowledge or to understand the biology of living systems; those who treat with chemical agents, disrupt, induce, knock -down or -out, interfere or in any other way manipulate living systems on the levels of organisms, tissues, cells and molecules; those who either experimentally or theoretically try to make sense of collected biological data, applying Bayesian, Boolean, fuzzy or just common logic; those who develop linear or non-linear models, studying or inferring relationships, classes, modules, functional categories, links, etc.; in summary all those who dare to attack multifaceted and hugely complex biological problems in a systematic way to submit papers to this section. The only requirements are the high quality, practical and theoretical relevance of the proposed research and association with the continually growing field of machine learning. Editors and reviewers of this section will endeavour to make our interactions as interesting for you as is possible through the World Wide Web.

January, 2006

Lev Soinov

Editor of the Bioinformatics Section,
Journal of Pattern Recognition Research

References

- [1] "The Sequence of the Human Genome," *Science*, The Celera Genomics Sequencing Team: Vol. 291, no. 5507, pp. 1304 - 1351, 2001.
- [2] "A physical map of the human genome," *Nature*, The International Human Genome Mapping Consortium: 409, pp. 934-941, 2001.
- [3] N.J. Mulder, et al., "InterPro, progress and status in 2005," *Nucleic Acids Res.* 33, Database Issue: D201-5, 2005.
- [4] H. Sahni, "REMBRANDT: Building a Robust Translational Research Framework for Brain Tumor Studies," *Bioinformatics Open Source Conference*, June, 2005.
- [5] J. Ihmels et al., "Revealing modular organization in the yeast transcriptional network," *Nat. Genet.*, 31, pp. 370-377, 2002.
- [6] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, 296, pp. 910-913, 2002.
- [7] J.A. Papin, J.L. Reed, B.O. Palsson, "Hierarchical thinking in network biology: the unbiased modularization of biochemical networks," *TRENDS in Biochemical Sciences*: Vol. 29, no. 12, 2004.
- [8] S.A. Kauffman, *The Origins of Order: Self Organization and Selection in Evolution*, Oxford University Press, Oxford, 1993.
- [9] T.R. Golub, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286, 531-537, 1999.
- [10] L.Dyrskjot, et al., "Identifying distinct classes of bladder carcinoma using microarrays," *Nat. Genet.*, 33, pp. 90-96, 2003.
- [11] E. Blaveri, et al., "Bladder cancer outcome and subtype classification by gene expression," *Clin Cancer Res.* 11(11):4044-55, 2005.
- [12] P. Pavlidis, et al., "Learning gene functional classifications from multiple data types," *J. Comput. Biol.*, 9, pp. 401-411, 2002.
- [13] A. Mateos, et al., "Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons," *Genome Res.*, 12, 1703-1715, 2002.

- [14] C.Z. Cai, et al., "SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Research*, 31(13):3692-7, 2003.
- [15] L.J. Jensen, et al., "Prediction of human protein function according to Gene Ontology categories," *Bioinformatics*, 19(5):635-42, 2003.
- [16] Bazzan, A.L., et al., "Automated annotation of keywords for proteins related to mycoplasmat-aceae using machine learning techniques," *Bioinformatics* 18 Suppl 2:S35-43, 2002.
- [17] C.H. Wu, et al. The Gene Ontology (GO) project in 2006. *Nucleic Acids Ressearch*, The Gene Ontology Consortium: 34: D322-D326, 2006.
- [18] C.H. Wu, "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Research*, 34: D187-D91, 2006.