Machine learning applications in bioinformatics

Jiří Kléma

Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27, Prague, Czech Republic
klema@labe.felk.cvut.cz,
WWW home page: http://labe.felk.cvut.cz/~klema/klema.html

Abstract. *Bioinformatics is a field of study dealing with methods for storing, retrieving and analyzing gene and protein oriented biological data. High-throughput technologies like DNA sequencing or microarrays allow researchers to obtain large volumes of heterogeneous and mutually interacting data. Analysis and understanding of these data provides a natural application field for machine learning algorithms. At the same time, bioinformatics is a scientific branch of such analytical complexity, data variety and abundance that it motivates further development of specialized learning algorithms such as co-clustering or multiple sequence alignment. This paper provides a brief overview of the topics and works discussed during my talk on machine learning applications in bioinformatics. The talk starts with a preview of fundamental bioinformatics analytical tasks solved by machine learning algorithms mentioning a few success stories. The second part summarizes the recent bioinformatics research carried out in my home research group, the Intelligent Data Analysis group of Czech Technical University.*

1 Analytical bioinformatics tasks

A complete overview of analytical bioinformatics tasks solvable and being solved by machine learning (ML) algorithms is out of scope of this short summary. [1] is a textbook that provides an introduction to the most important problems in computational biology and a unified treatment of the ML methods for solving these problems. The book is self-contained, its large part focuses on the principles of fundamental ML algorithms. A relevant concise review appeared in [2], its updated recent modification was presented in [3]. The reviews distinguish four principal classes of tasks. Firstly, a large group of bioinformatics problems can be posed as classification tasks. Genome annotation including gene finding and searching for DNA binding sites with proteins or gene function prediction and protein secondary structure prediction make examples. Secondly, clustering can be used to learn functional similarity from gene expression data or it can form phylogenetic trees. Thirdly, probabilistic graphical models can serve for modelling of DNA sequences in genomics or inference of genetic networks in systems biology. Last but not least, optimization algorithms have been proposed to solve the multiple sequence alignment problem or they appear in simplified models of protein folding.

1.1 Success stories and interactions

The bioinformatics tool with the largest impact is undoubtedly The Basic Alignment Search Tool (BLAST) and its successors [4] for searching a large sequence database against a query sequence. The NCBI server that provides the service with heuristic methods for sequence database searching handles more than half a million queries a day, the paper [4] introducing the improved PSI-BLAST has tens of thousands of citations. Another success story is an early case study on predictive classification from gene expression data [5]. The study proved feasibility of cancer classification based solely on gene expression monitoring. Although other latter studies showed that this positive result cannot be by means taken for granted, since then molecular classification is an option in disease diagnostics.

Bioinformatics directly motivates some cutting edge ML projects such as automated hypotheses generation and learning of optimal workflows. [6] reports the development of Robot Scientist “Adam”, which autonomously generated functional genomics hypotheses about the yeast Saccharomyces cerevisiae and experimentally tested these hypotheses by using laboratory automation. One of its main objectives of the ongoing European ML and data mining project e-LICO [7, 8] is to implement an intelligent data mining assistant that takes in user specifications of the learning task and the available data, plans a methodologically correct learning process, and suggests workflows that the user can execute to achieve the prespecified objectives. Bioinformatics is the major application area.

2 IDA bioinformatics research topics

One of our main research topics is learning from gene expression data driven by background knowledge [9]. Mining patterns from gene expression data represents an alternative way to clustering [10]. Clustering provides the most straightforward and traditional approach to obtain co-expressed genes. However, a typical group of genes shares an activation pattern only under specific experimental conditions. Local methods such as pattern mining can identify exactly the sets of genes displaying a specific expression characteristic in a set of situations. The main bottleneck of this type of analysis is twofold – computational costs and an overwhelming number of candidate patterns which can hardly be further exploited. A timely application of background knowledge available in literature databases, gene ontologies and other sources can help to focus on the most plausible patterns only. Molecular classification of biological samples based on their gene-expression profiles is a natural learning task with immediate practical uses. Nevertheless, molecular classifiers based solely on gene expression in most cases cannot be considered useful decision-making tools or decision-supporting tools. Similarly to the domain of pattern mining, recent efforts in the field of molecular classification aim to employ background knowledge. The idea is to extract features that correspond to functionally related gene sets instead of the individual genes, respectively the probesets whose expression is available in the original expression data [11, 12].

The previous paragraph employs the available structural genomic knowledge to improve the analysis of gene expression data. We also studied several methods to create it from collections of free biomedical texts, namely the research papers and their short summaries [13]. [14] proposes a novel ball-histogram approach to DNA-binding propensity prediction of proteins.

Last but not least, the IDA group cooperates with several biological institutes and labs. To exemplify, [15] shows an application of the set-level approach discussed above to the particular domain of respirable ambient air particulate matter, the principal research partner was the Department of Genetic Ecotoxicology from Czech Academy of Sciences. [16] evaluates differences in the intragraft transcriptome after successful induction therapy using two rabbit antithymocyte globulins, the partner was the Department of Nephrology, Transplant Center, Institute for Clinical and Experimental Medicine.

References

1. P. Baldi,S. Brunak: *Bioinformatics: the machine learning approach*, 2nd edition, MIT Press, 2001, 452.
2. P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafe, A. Perez, V. Robles: *Machine learning in bioinformatics.* Briefings in Bioinformatics, 7(1), 2005, 86–112.
3. I. Inza, B. Calvo, R. Armananzas, E. Bengoetxea, P. Larranaga, J. A. Lozano: *Machine learning: an indispensable tool in bioinformatics.* Methods Mol. Biol. 593, 2010, 25–48.
4. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman: *Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.* Nucleic Acids Research, 25, 1997, 3389–3402.
5. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. , J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander: *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 286 (5439), 1999, 531–537.
6. R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, A. Clare: *The automation of science.* Science 324 (5923), 2009, 85–89.
7. e-lico project: An e-laboratory for interdisciplinary collaborative research in data mining and data-intensive science, `http://www.e-lico.eu/`, August 2012.
8. M. Hilario, P. Nguyen, H. Do, A. Woznica, A. Kalousis: *Ontology-based meta-mining of knowledge discovery workflows.* In Jankowski, N., Duchs, W. Grabczewski, K., Meta-Learning in Computational Intelligence, Springer, 2011, 273–316.
9. J. Klema: *Learning from heterogeneous genomic data.* FEE CTU, habilitation thesis, to appear.
10. J. Klema, S. Blachon, A. Soulet, B Cremilleux, O. Gandrilon: *Constraint-based knowledge discovery from SAGE data.* In Silico Biology, 8, 0014, 2008.
11. M. Holec, J. Klema, F. Zelezny, J. Tolar: *Comparative evaluation of set-level techniques in predictive classification of gene expression samples.* BMC Bioinformatics, 13, (10), 2012, S15.
12. M. Krejnik, J. Klema: *Empirical evidence of the applicability of functional clustering through gene expression classification.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(3), 2012, 788–798.
13. M. Plantevit, T. Charnois, J. Klema, C. Rigotti, B. Cremilleux: *Combining sequence and itemset mining to discover named entities in biomedical texts: A new type of pattern.* International Journal of Data Mining, Modelling and Management, 1(2), 2009, 119–148.
14. A. Szaboova, O. Kuzelka, F. Zelezny, J. Tolar: *Prediction of DNA-binding propensity of proteins by the ball-histogram method using automatic template search BMC.* Bioinformatics 13 (10), 2012, 3.
15. H. Libalova, K. Uhlirova, J. Klema, M. Machala, R. Sram, M. Ciganek, J. Topinka: *Global gene expression changes in human embryonic lung fibroblasts induced by organic extracts from respirable air particles.* Particle and Fibre Toxicology, 9(1), 2012.
16. M. Urbanova, I. Brabcova, E. Girmanova, F. Zelezny, O. Viklicky: *Differential regulation of the nuclear factor-kappa-B pathway by rabbit antithymocyte globulins in kidney transplantation.* Transplantation 93(6), 2012, 589–96.