Machine Learning, 52, 5–9, 2003

Machine Learning in the Genomics Era Editorial: Methods in Functional Genomics

PAOLA SEBASTIANI sebas@math.umass.edu
Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA, USA

ISAAC S. KOHANE isaac_kohane@harvard.edu
Children's Hospital Informatics Program, Harvard Medical School, Boston, MA, USA

MARCO F. RAMONI marco_ramoni@harvard.edu
Children's Hospital Informatics Program, Harvard Medical School, Boston, MA, USA

1. Introduction

In June 2000, leaders of the Human Genome Project, Craig Venter of Celera Genomics, and U.S. President Clinton announced the completion of a "working draft" of the human genome: the genetic blueprint of a human being. Today, the legacy of that announcement is the challenge to annotate this map, by understanding the roles and functions of genes—and their interplay with proteins and the environment—to create complex, dynamic living systems. This understanding is the goal of functional genomics.

Functional genomics has recently become a major focus of machine learning applications thanks to the development of the new technology of DNA or expression microarray (Schena et al., 1995; Lockhart et al., 1996). Microarrays enable investigators to observe the genome of entire organisms in action by simultaneously measuring the level of activation of thousands of genes under the same experimental conditions. This technology provides today unprecedented discovery opportunities and is reshaping biomedical sciences by shifting its paradigm from a hypothesis driven to a data driven approach (Lander, 1999). Not surprisingly, parallel to these technological advances has been the development of machine learning methods able to integrate and understand the data generated by this new kind of experiments. However, most of this research has been conducted outside the traditional machine learning research community. The aim of this special issue is to bridge this divide by reporting methodological advances in automated learning from functional genomic data to the core machine learning community.

2. Microarray technology

The modern concept of gene expression dates back to the seminal work of Jacob and Monod (1961) and their fundamental discovery that differential gene expression—that is when and in what quantities a gene is expressed—determines differential protein abundance that

induces different cell functions (note that this one-to-one correspondence is not necessarily true even though it is a useful heuristic or generalization). During its expression, a gene transcribes its DNA sequence into molecules of mRNA (messenger Ribonucleic acid) that are then transported out of the cell nucleus and used as a template for making a protein. This two-step representation of the protein-synthesis process constitutes the *central dogma of molecular biology* (Crick, 1970).

Since the first step of a gene expression consists of copying its DNA code into mRNA molecules, the amount of mRNA molecules produced during this process provides a measure of the gene expression level. The basic idea behind microarray technology is to measure the expression level of thousands of genes in a particular cell or collection of cells by simultaneously measuring their mRNA abundance. Technically, a microarray is a platform gridded in such a way that each location of the grid corresponds to a gene and contains several copies of the DNA sequence of the gene (cDNA microarrays) or several copies of a short specific segments—known as *synthetic oligonucleotides*—characteristic of the gene (synthetic oligonucleotide microarrays) (Duggan et al., 1999). The tethered copies of DNA sequences or synthetic oligonucleotides are called the *probes*. To measure the relative expression level of the genes in a particular cell, investigators prepare the *target* by extracting the mRNA from the cell and making a fluorescence-tagged copy of this mRNA. This tagged copy is then hybridized to the probes in the microarray. During this process, if a gene is expressed in the target cells, its mRNA representation will bind to the probes on the microarray, and its fluorescence tagging will make the corresponding probe on the microarray brighter. Studies have demonstrated that the brightness of a probe is correlated with the amount of mRNA in the original sample and, therefore, to the expression level of each gene.

Aside from some technical differences—described for example in Kohane, Kho, and Butte (2002) and Sebastiani, Kohane, and Ramoni (2003)—both cDNA and oligonucleotide microarrays provide a panoramic view of the activity of genes under particular experimental conditions, and are nowadays used to answer the same broad classes of questions.

3. Analysis of microarray experiments

Typical experimental questions investigated by microarray experiments are the detection of genes differentially expressed across two different experimental conditions, such as normal and tumor tissues. Microarray experiments can answer a variety of questions: they can be used to build the expression profile of a particular tissue, say breast cancer tissue, and develop new diagnostic devices; they can help to identify new classes of cells; they can elucidate the control mechanisms underlying the expression dynamics of the genome.

One of the main analytical challenges of microarray experiments is that the technology is still comparatively expensive and that some cell types, such as some kinds of tumors, are relatively rare or difficult to acquire. Investigators are therefore usually faced with the task of identifying differences, discovering new classes or decoding control mechanisms in thousands of genes with a very small set of experiments.

In the simplest experimental setting, comparing the molecular profile of cells in two experimental conditions, differential analysis is carried out to select the genes that have substantial differential expression in the two conditions. The aim of these experiments is

to build classifiers able to diagnose molecular differences hard, or sometimes impossible, to identify using traditional methods. Popular techniques are based on the empirical fold change—the ratio of the expression sample means in the two conditions—or some standardization of the sample mean difference, and threshold values are often computed by using permutation tests (Golub et al., 1999; Tusher, Tibshirani, & Chu, 2000). To investigate whether the selected genes are predictive of the class, classification models are often built from the expression data of the selected genes. Typical approaches are based on nearest neighbor classification (Golub et al., 1999), support vector machines, or other discriminant analysis techniques (Dudoit, Fridlyand, & Speed, 2002).

The determination of the error rate of these classifiers using a relatively small number of samples is a critical problem to foster the development and the application of these new diagnostic methods. Hsing et al. in their contribution *Relation between permutation test p-values and classifier error estimates* examine this important issue, and the validity of permutation procedures to assess the dependency of the error rate from the particular data set used to build the classifier. Particularly, they show that random labeling does not provide any further insight into the accuracy of the classifier or the precision of the error estimate. Relevant to the problem of building classification models from gene expression data is the contribution by Long and Vega *Boosting and microarray data*. The authors identify situations in which boosting does not perform well, particularly when a strong association exists between the gene expression profile and the class designation, and identify modifications to improve the construction of class prediction rules.

When the objective of the microarray data analysis is to group genes with a similar expression profile, or samples with a similar molecular profile, investigators use unsupervised classification techniques, such as clustering or multidimensional scaling. Given the exploratory nature of these studies and the multi-dimensionality of the data, it is critical to the success of an application to be able to graphically display the results in a manner amenable to biologists. Very popular techniques for clustering gene expression profiles or sample molecular profiles are hierarchical clustering (Eisen et al., 1998) and self-organizing maps (Kohonen, 1997). Self-organizing maps have been used, for instance, to identify classes of genes with similar functions in the yeast cell cycle (Tamayo et al., 1999) and they have been combined with the nearest neighbor classification method to discriminate between different types of acute leukemia (Golub et al., 1999). The contribution of Hautaniemi et al. *Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps* addresses the issue of visualization of the results. Their study shows that self-organizing maps offer excellent resources for visualization, analysis, and interpretation of the vast amounts of multi-dimensional gene expression data. Although self-organizing maps are a form of k-mean clustering, tailored to atemporal data, they are often used to group gene expression profiles from temporal experiments. Zhang et al. in their contribution *Self-organizing latent lattice models for temporal gene expression profiling* develop a learning method designed for temporal gene expression profiling. Their method learns probabilistic lattice maps of the gene expressions, which are then used for profiling the trajectories of temporal expressions of co-regulated genes. This self-organizing latent lattice (SOLL) model combines the topographic mapping capability of self-organizing maps and the generative property of probabilistic latent-variable models.

A critical issue of any clustering method is the determination of the cardinality of the cluster set. Monti et al., in their contribution *Consensus clustering*, present a new method of class discovery and clustering validation that addresses both the issues of the dependency of the clusters on the specific data set, and the determination of the optimal number of clusters. Their method, called *consensus clustering* uses the agreement across multiple runs of a clustering algorithm to assess the stability of the discovered clusters to random perturbations of the data. They show that this method can also be used to represent the consensus over multiple runs of a clustering algorithm with random restart to inspect several features of the clustering model, including the best number of clusters and the robustness of cluster membership.

One of the intuitions behind the use of standard machine learning techniques for clustering gene expression profiles is that genes belonging to the same cluster have similar functions (Eisen et al., 1998). Provided this intuition is correct, it is important to develop automated ways to integrate standard machine learning techniques with the increasing information available about particular genes. Raychaudhuri et al. address this issue in their contribution *Inclusion of textual documentation in the analysis of multidimensional data sets: application to gene expression data*. The authors present a method that integrate text analysis with gene expression data analysis to find meaningful gene expression patterns that correlate with the biology described in scientific literature. Their method—called *neighbor divergence per gene*—assigns a score to a given subgroup of genes indicating the likelihood that the genes share a biological property or function and an algorithm that searches for linear projections in gene expression data that separate functionally related groups of genes from the other genes.

In microarray studies, clustering techniques are also used for comparative analysis of gene expression data collected in a variety of conditions. However, clustering techniques by themselves are unable to account for the dependency structure underlying the functions of genes. Boolean networks are currently used as knowledge representation formalisms to capture these interactions among genes. A critical step to identify these *regulatory networks* from gene expression data is to find a network that is consistent with the given observations or determine whether such a network exists at all. This problem—known as the Consistency Problem—is considered in *On Learning Gene Regulatory Networks Under the Boolean Network Model* by Lädesmäki et al., which addresses the specific issue that gene expression data are not noise-free and present algorithms to find one or all Boolean networks relatively consistent with a set of the gene expression measurements.

Besides their important role as knowledge representation formalism, Boolean networks can help investigators understand the co-regulation of genes and the behavior of a biology system under particular controllable conditions. This problem is examined in *External control in Markovian genetic regulatory networks* by Datta et al. They present a procedure to identify the sequence of control actions that minimize some performance index in a finite number of steps, thus gaining the ability to identify when to suspend treatments in patients and observe the consequences before making the decision as to whether further intervention is necessary.

Eric Lander (Lander, 1999), one of the most influential scientists of the genomic era and a mathematician by training and profession, wrote that developing experimental designs able to take advantage of the full power of microarray technology is the challenge for biologists

of this century but he also acknowledged that, to fully understand the results of microarray experiments, new analytical perspectives are needed. The contributions to this special issue tackle some fundamental issues in the application of machine learning methods to functional genomics and, in so doing, provide an outlook on the variety of computational challenges laying ahead.

We are proud to present the machine learning community with this *randonnee* at the heart of the exciting area at the intersection between artificial intelligence and biological sciences, and we are grateful to the authors and the referees for creating it.

Acknowledgment

P. Sebastiani was partially supported by the NSF program in Bioengineering and Environmental Systems Division/Biotechnology under ContractECS-0120309. I.S. Kohane was partially supported by the NHLBI Program in Genomic Applications ŞGenomics of Cardiovascular Development, Adaptation, and RemodelingT̂ U01 HL066582. M.F. Ramoni was partially supported by the NSF program in Bioengineering and Environmental Systems Division/Biotechnology under Contract ECS-0120309.

References

Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature, 227*, 561–563.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97:576*, 77–87.

Duggan, J. D., Bittner, M., Chen, Y., Meltzer, P., & Trent, J. M. (1999). Expression profiling using CDNA microarrays. *Nature Genetics Supplement, 21*, 10–14.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA, 95*, 14863–14868.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 15*, 531–537.

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol., 3*, 318–356.

Kohane, I. S., Kho, A., & Butte, A. J. (2002). *Microarrays for an Integrative Genomics*. Cambridge, MA: MIT Press.

Kohonen, T. (1997). *Self Organizing Maps*. Berlin, DE: Springer.

Lander, E. S. (1999). Array of hope. *Nature Genetics Supplement, 21*, 3–4.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., & Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology, 14*, 1675–1680.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270:5235*, 467–470.

Sebastiani, P., Kohane, E. G. I., & Ramoni, M. (2003). Statistical challenges in functional genomics (with discussion). *Statistical Science*, to appear.

Tamayo, P., Slonim, D., Mesirov, J., Zhui, Q., Kitareewani, S., Dmitrovsky, E., Lander, E. S., & Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA, 96*, 2907–2912.

Tusher, V. G., Tibshirani, R., & Chu, G. (2000). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA, 98*, 5116–5121.