# MACHINE LEARNING IN COMPUTATIONAL BIOLOGY

Cornelia Caragea and Vasant Honavar
Department of Computer Science
Iowa State University
cornelia@cs.iastate.edu, honavar@cs.iastate.edu

## SYNONYMS

Data Mining in Computational Biology; Data Mining in Bioinformatics; Machine Learning in Bioinformatics; Machine Learning in Systems Biology; Data Mining in Systems Biology

## DEFINITION

Advances in high throughput sequencing and "omics" technologies and the resulting exponential growth in the amount of macromolecular sequence, structure, gene expression measurements, have unleashed a transformation of biology from a data-poor science into an increasingly data-rich science. Despite these advances, biology today, much like physics was before Newton and Leibnitz, has remained a largely descriptive science. Machine learning [6] currently offers some of the most cost-effective tools for building predictive models from biological data, e.g., for annotating new genomic sequences, for predicting macromolecular function, for identifying functionally important sites in proteins, for identifying genetic markers of diseases, and for discovering the networks of genetic interactions that orchestrate important biological processes [3]. Advances in machine learning e.g., improved methods for learning from highly unbalanced datasets, for learning complex structures of class labels (e.g., labels linked by directed acyclic graphs as opposed to one of several mutually exclusive labels) from richly structured data such as macromolecular sequences, 3-dimensional molecular structures, and reliable methods for assessing the performance of the resulting models, are critical to the transformation of biology from a descriptive science into a predictive science.

## HISTORICAL BACKGROUND

Large scale genome sequencing efforts have resulted in the availability of hundreds of complete genome sequences. More importantly, the GenBank repository of nucleic acid sequences is doubling in size every 18 months [4]. Similarly, structural genomics efforts have led to a corresponding increase in the number of macromolecular (e.g., protein) structures [5]. At present, there are over a thousand databases of interest to biologists [16]. The emergence of high-throughput "omics" techniques, e.g., for measuring the expression of thousands of genes under different perturbations, has made possible system-wide measurements of biological variables [8]. Consequently, discoveries in biological sciences are increasingly enabled by machine learning.

Some representative applications of machine learning in computational and systems biology include: Identifying the protein-coding genes (including gene boundaries, intron-exon structure) from genomic DNA sequences; Predicting the function(s) of a protein from its primary (amino acid) sequence (and when available, structure and its interacting partners); Identifying functionally important sites (e.g., protein-protein, protein-DNA, protein-RNA binding sites, post-translational modification sites) from the protein's amino acid sequence and, when available, from the protein's structure; Classifying protein sequences (and structures) into structural classes; Identifying functional modules (subsets of genes that function together) and genetic networks from gene

This work was funded in part by grants from the National Institutes of Health (GM 066387) and the National Science Foundation (IIS 0711356) to Vasant Honavar

expression data.

These applications collectively span the entire spectrum of machine learning problems including supervised learning, unsupervised learning (or cluster analysis), and system identification. For example, protein function prediction can be formulated as a supervised learning problem: given a dataset of protein sequences with experimentally determined function labels, induce a classifier that correctly labels a novel protein sequence. The problem of identifying functional modules from gene expression data can be formulated as an unsupervised learning problem: given expression measurements of a set of genes under different conditions (e.g., perturbations, time points), and a distance metric for measuring the similarity or distance between expression profiles of a pair of genes, identify clusters of genes that are co-expressed (and hence are likely to be co-regulated). The problem of constructing gene networks from gene expression data can be formulated as a system identification problem: given expression measurements of a set of genes under different conditions (e.g., perturbations, time points), and available background knowledge or assumptions, construct a model (e.g., a boolean network, a bayesian network) that explains the observed gene expression measurements and predicts the effects of experimental perturbations (e.g., gene knockouts).

## SCIENTIFIC FUNDAMENTALS

Challenges presented by computational and systems biology applications have driven, and in turn benefited from, advances in machine learning. We proceed to describe some of these developments below.

**Multi-label classification:** In the traditional classification problem, an instance $\mathbf{x}_i$, $i = 1, \cdots, n$, is associated with a single class label $y_j$ from a finite, disjoint set of class labels $Y$, $j = 1, \cdots, k$, $k = |Y|$ (*single-label classification problem*). If the set $Y$ has only two elements, then the problem is refered to as the *binary classification problem*, otherwise, if $Y$ has more than two elements, then it is refered to as *multi-class classification problem*. However, in many biological applications, an instance $\mathbf{x}_i$ is associated with a subset of, not necessarily disjoint, class labels in $Y$ (*multi-label classification problem*). For example, many genes and proteins are multi-functional. Most of the existing algorithms cannot simultaneously label a gene or protein with several, not necessarily mutually exclusive functions. Each instance is then assigned to a subset of nodes in the hierarchy, yielding a *hierarchical multi-label classification problem* or a *structured output classification problem*. The most common approach to dealing with *multi-label classification problem* [7] is to transform the problem into $k$ binary classification problems, one for each different label $y_j \in Y$, $j = 1, \cdots, k$. The transformation consists of constructing $k$ datasets, $D_j$, each containing all instances of the original dataset, such that an instance in $D_j$, $j = 1, \cdots, k$, is labeled with 1 if it has label $y_j$ in the original dataset, and 0 otherwise. During classification, for a new unlabeled instance $\mathbf{x}_{test}$, each individual classifier $C_j, j = 1, \cdots, k$, returns a prediction that $\mathbf{x}_{test}$ belongs to the class label $y_j$ or not. However, the transformed datasets that result from this approach are highly unbalanced, typically, with the number of positively labeled instances being significantly smaller than the number of negatively labeled instances, requiring the use of methods that can cope with unbalanced data. Alternative evaluation metrics need to be developed for assessing the performance of multi-label classifiers. This task is complicated by correlations among the class labels.

**Learning from unbalanced data:** Many of the macromolecular sequence classification problems present the problem of learning from highly *unbalanced* data. For example, only a small fraction of amino acids in an RNA-binding protein binds to RNAs. Classifiers that are trained to optimize accuracy generally perform rather poorly on the minority class. Hence, if accurate classification of instances from the minority class is important (or equivalently, the false positives and false negatives have unequal costs or risks associated with them), it is necessary to change the distribution of positive and negative instances *during training* by randomly selecting a subset of the training data for the majority class, or alternatively, assigning different *weights* to positive and negative samples (and learn from the resulting weighted samples). More recently, *ensemble classifiers* [12] have been shown to improve the performance of sequence classifiers on unbalanced datasets. Unbalanced datasets also complicate both the training and the assessment of the predictive performance of classifiers. *Accuracy* is not a useful performance measure in such scenarios. Indeed, no single performance measure provides a complete picture of the classifier's performance. Hence, it is much more useful to examine ROC (Receiver Operating Characteristic) or precision-recall curves [3]. Of particular interest are methods that can directly optimize

alternative performance measures that take into account the unbalanced nature of the dataset and user-specified tradeoff between false positive and false negative rates.

**Data representation:** Many computational and systems biology applications of machine learning present challenges in data representation. Consider for example, the problem of identifying functionally important sites (e.g., RNA-binding residues) from amino acid sequences. In this case, given an amino acid sequence, the classifier needs to assign a binary label (1 for an RNA-binding residue and 0 for a non RNA-binding residue) to each letter of the sequence. To solve this problem using standard machine learning algorithms that work with a fixed number of input features, it is fairly common to use a *sliding window* approach [11] to generate a collection of fixed length windows, where each window corresponds to the target amino acid and an equal number of its sequence neighbors on each side. The classifier is trained to label the target residue. Similarly, identifying binding sites from a 3-dimensional structure of the protein requires transforming the problem into one that can be handled by a traditional machine learning method. Such transformations, while they allow the use of existing machine learning methods on macromolecular sequence and structure labeling problems, complicate the task of assessing the performance of the resulting classifier (see below).

**Performance Assessment:** Standard approaches to assessing the performance of classifiers rely on $k$-fold cross-validation wherein a dataset is partitioned into $k$ disjoint subsets (folds). The performance measure of interest is estimated by averaging the measured performance of the classifier on $k$ runs of a cross-validation experiment, each using a different choice of the $k-1$ subsets for training and the remaining subset for testing the classifier. The fixed length window representation described above complicates this procedure on macromolecular sequence labeling problems: The training and test sets obtained by random partitioning of the dataset of labeled windows can contain windows that originate from the same sequence, thereby violating a critical requirement for cross-validation, namely, that the training and test data be disjoint. The resulting overlap between training and test data can yield overly optimistic estimates of performance of the classifier. A better alternative is to perform sequence-based (as opposed to window-based) cross-validation by partitioning the set of sequences (instead of windows) into disjoint folds. This procedure guarantees that training and test sets are indeed disjoint [9]. Obtaining realistic estimates of performance in sequence classification and sequence labeling problems also requires the use of *non-redundant* datasets [14].

**Learning from sparse datasets:** In gene expression datasets the number of genes is typically in the hundreds or thousands whereas the number of measurements (conditions, perturbations) is typically fewer than ten. This presents significant challenges in infering genetic network models from gene expression data because the number of variables (genes) far exceeds the number of observations or data samples. Approaches to dealing with this challenge require reducing the effective number of variables via variable selection [17] or abstraction i.e., by grouping variables into clusters that behave similarly under the observed conditions. Another approach to dealing with sparsity of data in such settings is to incorporate information from multiple datasets [18].

## KEY APPLICATIONS*

**Protein function prediction:** Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters and molecular machines in cells. Understanding protein function is critical to understanding diseases and ultimately in designing new drugs. Until recently, the primary source of information about protein function has come from biochemical, structural, or genetic experiments on individual proteins. However, with the rapid increase in number of genome sequences, and the corresponding growth in the number of protein sequences, the numbers of experimentally determined structures and functional annotations has significantly lagged the number of protein sequences. With the availability of datasets of protein sequences with experimentally determined functions, there is increasing use of sequence or structural homology based transfer of annotation from already annotated sequences to new protein sequences. However, the effectiveness of such homology-based methods drops dramatically when the sequence similarity between the target sequence and the reference sequence falls below 30%. In many instances, the function of a protein is determined by conserved local sequence sequence motifs. However, approaches that assign function to a protein based on the presence of a single motif (the so-called characteristic motif) fail to take advantage of multiple sequence motifs that are

correlated with critical structural features (e.g., binding pockets) that play a critical role in protein function. Against this background, machine learning methods offer an attractive approach to training classifiers to assign putative functions to protein sequences. Machine learning methods have been applied, with varying degrees of success, to the problem of protein function prediction. Several studies have demonstrated that machine learning methods, used in conjunction with traditional sequence or structural homology based techniques and sequence motif-based methods outperform the latter in terms of accuracy of function prediction (based on cross-validation experiments). However, the efficacy of alternative approaches in genome-wide prediction of functions of protein-coding sequences from newly sequenced genomes remains to be established. There is also significant room for improving current methods for protein function prediction.

**Identification of potential functional annotation errors in genes and proteins:** As noted above, to close the sequence-function gap, there is an increasing reliance on automated methods in large-scale genome-wide annotation efforts. Such efforts often rely on transfer of annotations from previously annotated proteins, based on sequence or structural similarity. Consequently, they are susceptible to several sources of error including errors in the original annotations from which new annotations are inferred, errors in the algorithms, bugs in the software used to process the data, and clerical errors on the part of human curators. The effect of such errors can be magnified because they can propagate from one set of annotated sequences to another. Because of the increasing reliance of biologists on reliable functional annotations for formulation of hypotheses, design of experiments, and interpretation of results, incorrect annotations can lead to wasted effort and erroneous conclusions. Hence, there is an urgent need for computational methods for checking consistency of such annotations against independent sources of evidence and detecting potential annotation errors. A recent study has demonstrated the usefulness of machine learning methods to *identify and correct* potential annotation errors [1].

**Identification of functionally important sites in proteins:** Protein-protein, protein-DNA, and protein-RNA interactions play a pivotal role in protein function. Reliable identification of such interaction sites from protein sequences has broad applications ranging from rational drug design to the analysis of metabolic and signal transduction networks. Experimental detection of interaction sites must come from determination of the structure of protein-protein, protein-DNA and protein-RNA complexes. However, experimental determination of such complexes lags far behind the number of known protein sequences. Hence, there is a need for development of reliable computational methods for identifying functionally important sites from a protein sequence (and when available, its structure, but not the complex). This problem can be formulated as a sequence (or structure) labeling problem. Several groups have developed and applied, with varying degrees of success, machine learning methods for identification of functionally important sites in proteins (see [21, 13, 22] for some examples). However, there is significant room for improving such methods.

**Discovery and analysis of gene and protein networks:** Understanding how the parts of biological systems (e.g., genes, proteins, metabolites) work together to form dynamic functional units, e.g., how genetic interactions and environmental factors orchestrate development, aging, and response to disease, is one of the major foci of the rapidly emerging field of systems biology [8]. Some of the key challenges include the following: uncovering the biophysical basis and essential macromolecular sequence and structural features of macromolecular interactions; comprehending how temporal and spatial clusters of genes, proteins, and signaling agents correspond to genetic, developmental and regulatory networks [10]; discovering topological and other characteristics of these networks [19]; and explaining the emergence of systems-level properties of networks from the interactions among their parts. Machine learning methods have been developed and applied, with varying degrees of success, in learning predictive models including boolean networks [20] and bayesian networks [15] from gene expression data. However, there is significant room for improving the accuracy and robustness of such algorithms by taking advantage of multiple types of data and by using active learning.

## FUTURE DIRECTIONS

Although many machine learning algorithms have had significant success in computational biology, several challenges remain. These include the development of: efficient algorithms for learning predictive models from distributed data; cumulative learning algorithms that can efficiently update a learned model to accomodate

changes in the underlying data used to train the model; effective methods for learning from sparse, noisy, high-dimensional data; and effective approaches to make use of the large amounts of unlabeled or partially labeled data; algorithms for learning predictive models from disparate types of data: macromolecular sequence, structure, expression, interaction, and dynamics; and algorithms that leverage optimal experiment design with active learning in settings where data is expensive to obtain.

## CROSS REFERENCE*

Biostatistics and data analysis; Biological Networks; Classification; Clustering; Graph Mining, Data Mining.

## RECOMMENDED READING

**Between 5 and 15 citations to important literature, e.g., in journals, conference proceedings, and websites.**


[1] C. Andorf, D. Dobbs, and V. Honavar. Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach. *BMC Bioinformatics*, doi:10.1186/1471-2105-8-284, 2007.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[3] P. Baldi and S. Brunak. *Bioinformatics: the Machine Learning Approach*. MIT Press, 2001.

[4] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. Genbank. *Nucleic Acids Research*, 35(Database issue):D21–D25, 2007.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning*. Berlin: Springer, 2006.

[7] M.R. Boutell, J. Luo, X. Shen, and C.M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37:1757–1771, 2004.

[8] F. J. Bruggeman and H.V. Westerhoff. The nature of systems biology. *Trends in Microbiology*, 15:45–50, 2007.

[9] C. Caragea, J. Sinapov, D. Dobbs, and V. Honavar. Assessing the performance of macromolecular sequence classifiers. In *IEEE 7th International Symposium on Bioinformatics and Bioengineering*, 320-326, IEEE Press, 2007.

[10] H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9:67–103, 2002.

[11] T. G. Diettrich. Machine learning for sequential data: A review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, 15–30, Springer-Verlag, 2002.

[12] T.G. Diettrich. Ensemble methods in machine learning. Springer-Verlag *Lecture Notes in Computer Science*, 1857:1–15, 2000.

[13] Y. El-Manzalawy, D. Dobbs, V. Honavar. Predicting linear B-cell epitopes using string kernels. *Journal of Molecular Recognition*, 21:243–255, 2008.

[14] Y. El-Manzalawy, D. Dobbs, V. Honavar. On evaluating MHC-II binding peptide prediction methods, *PLoS One*, In press, 2008

[15] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[16] M. Y. Galperin. The molecular biology database collection: 2008 update. *Nucleic Acids Research*, 36:D2–4, 2008.

[17] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[18] L. Hecker, T. Alcon, V. Honavar, and H. Greenlee. Querying multiple large-scale gene expression datasets from the developing retina using a seed network to prioritize experimental targets. *Bioinformatics and Biology Insights*, 2:91–102, 2008.

[19] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[20] H. Lahdesmaki, I. Shmulevich, and O. Yli-Harja. On learning gene regulatory networks under the boolean network model. *Machine Learning*, 52:147–167, 2007.

[21] M. Terribilini, J.-H. Lee, C. Yan, R. L. Jernigan, V. Honavar, and D Dobbs. Predicting RNA-binding sites from amino acid sequence. *RNA Journal*, 12:1450-1462, 2006.

[22] C. Yan, M. Terribilini, F. Wu, R.L. Jernigan, D. Dobbs, and V. Honavar. Identifying amino acid residues involved in protein-DNA interactions from sequence. *BMC Bioinformatics*, doi:10.1186/1471-2105-7-262, 2006.