MACHINE LEARNING
IN BIOINFORMATICS
Wiley Series on
Bioinformatics: Computational Techniques and Engineering

Bioinformatics and computational biology involve the comprehensive application of mathematics, statistics, science, and computer science to the understanding of living systems. Research and development in these areas require cooperation among specialists from the fields of biology, computer science, mathematics, statistics, physics, and related sciences. The objective of this book series is to provide timely treatments of the different aspects of bioinformatics spanning theory, new and established techniques, technologies and tools, and application domains. This series emphasizes algorithmic, mathematical, statistical, and computational methods that are central in bioinformatics and computational biology.

Series Editors: Professor Yi Pan and Professor Albert Y. Zomaya
pan@cs.gsu.edu  zomaya@it.usyd.edu.au

Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications
Xiaohua Hu and Yi Pan

Grid Computing for Bioinformatics and Computational Biology
Edited by El-Ghazali Talbi and Albert Y. Zomaya

Bioinformatics Algorithms: Techniques and Applications
Ion Mandiou and Alexander Zelikovsky

Analysis of Biological Networks
Edited by Björn H. Junker and Falk Schreiber

Machine Learning in Bioinformatics
Edited by Yan-Qing Zhang and Jagath C. Rajapakse
CONTENTS

Foreword ix
Preface xi
Contributors xvii

1 Feature Selection for Genomic and Proteomic Data Mining 1
   Sun-Yuan Kung and Man-Wai Mak

2 Comparing and Visualizing Gene Selection and Classification Methods for Microarray Data 47
   Rajiv S. Menjoge and Roy E. Welsch

3 Adaptive Kernel Classifiers Via Matrix Decomposition Updating for Biological Data Analysis 69
   Hyunsoo Kim and Haesun Park

4 Bootstrapping Consistency Method for Optimal Gene Selection from Microarray Gene Expression Data for Classification Problems 89
   Shaoning Pang, Ilkka Havukkala, Yingjie Hu, and Nikola Kasabov

5 Fuzzy Gene Mining: A Fuzzy-Based Framework for Cancer Microarray Data Analysis 111
   Zhenyu Wang and Vasile Palade

6 Feature Selection for Ensemble Learning and Its Application 135
   Guo-Zheng Li and Jack Y. Yang
CONTENTS

7 Sequence-Based Prediction of Residue-Level Properties in Proteins 157
Shandar Ahmad, Yemlembam Hemjit Singh, Marcos J. Araúzo-Bravo, and Akinori Sarai

8 Consensus Approaches to Protein Structure Prediction 189
Dongbo Bu, ShuaiCheng Li, Xin Gao, Libo Yu, Jinbo Xu, and Ming Li

9 Kernel Methods in Protein Structure Prediction 209
Jayavardhana Gubbi, Alistair Shilton, and Marimuthu Palaniswami

10 Evolutionary Granular Kernel Trees for Protein Subcellular Location Prediction 229
Bo Jin and Yan-Qing Zhang

11 Probabilistic Models for Long-Range Features in Biosequences 241
Li Liao

12 Neighborhood Profile Search for Motif Refinement 263
Chandan K. Reddy, Yao-Chung Weng, and Hsiao-Dong Chiang

13 Markov/Neural Model for Eukaryotic Promoter Recognition 283
Jagath C. Rajapakse and Sy Loi Ho

14 Eukaryotic Promoter Detection Based on Word and Sequence Feature Selection and Combination 301
Xudong Xie, Shuanhu Wu, and Hong Yan

15 Feature Characterization and Testing of Bidirectional Promoters in the Human Genome—Significance and Applications in Human Genome Research 321
Mary Q. Yang, David C. King, and Laura L. Elnitski

16 Supervised Learning Methods for MicroRNA Studies 339
Byoung-Tak Zhang and Jin-Wu Nam

17 Machine Learning for Computational Haplotype Analysis 367
Phil H. Lee and Hagit Shatkay
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Machine Learning Applications in SNP–Disease Association Study</td>
<td>389</td>
</tr>
<tr>
<td></td>
<td><em>Pritam Chanda, Aidong Zhang, and Murali Ramanathan</em></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Nanopore Cheminformatics-Based Studies of Individual Molecular</td>
<td>413</td>
</tr>
<tr>
<td></td>
<td>Interactions</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Stephen Winters-Hilt</em></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>An Information Fusion Framework for Biomedical Informatics</td>
<td>431</td>
</tr>
<tr>
<td></td>
<td>*Srivatsava R. Ganta, Anand Narasimhamurthy, Jyotsna Kasturi, and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Raj Acharya</td>
<td></td>
</tr>
</tbody>
</table>

**Index** 453
Machine learning is a subfield of artificial intelligence and is concerned with the development of algorithms and techniques that allow computers to learn. It has a wide spectrum of applications such as natural language processing, search engines, medical diagnosis, bioinformatics and cheminformatics, stock market analysis, computer vision, and game playing. Recently, the amount of biological data requiring analysis has exploded and many machine learning methods have been developed to deal with this explosion of data. Hence, machine learning in bioinformatics has become an important research area for both computer scientists and biologists.

The aim of this book is to provide applications of machine learning to problems in the biological sciences, with particular emphasis on problems in bioinformatics. The book consists of a number of stand-alone chapters that explain and apply machine learning methods to central bioinformatics problems such as feature selection, sequence-based prediction of residue-level properties, promoter recognition, protein structure prediction, gene selection, and SNPS selection, classification, and data mining. This book represents the unification of two important fields in sciences—biology and computer science—with machine learning as a common theme. The chapters are written by well-known researchers in these interdisciplinary areas, and applications of various machine learning methods to different bioinformatics problems are presented. Students and scientists in biology and computer science will find this book valuable and accessible.

Several books in the similar areas have been published. However, this book is unique in that it presents cutting-edge research topics and methodologies in the area of machine learning methods when applied to bioinformatics. Many results presented in this book have never been published in the literature and represent the most advanced technologies in this exciting area. It also provides a comprehensive and balanced blend of topics, implementations, and case studies. I firmly believe that this book will further facilitate collaboration between machine learning researchers and bioinformaticians.

Both editors, Dr. Yan-Qing Zhang and Dr. Jagath C. Rajapakse, are rising stars in the areas of machine learning and bioinformatics. They have achieved a lot of research results in these areas. Their vision of creating such a book in a timely manner deserves our loud applause. This book is ideally suited both as a reference and as a text for a graduate course on machine learning or bioinformatics. This book can also serve as a
repository of significant reference materials because the references cited in each chapter serve as useful sources for further study in this area.

I highly recommend this timely and valuable book. I believe that it will benefit many readers and contribute to the further development of machine learning in bioinformatics.

Atlanta, Georgia
August 2008

Dr. Yi Pan
Chair and Professor
Georgia State University
In recent decades, machine learning techniques have been widely applied to bioinformatics. Many positive results have indicated that machine learning methods are useful for solving complex biomedical problems too difficult to solve by experts. Traditionally, researchers do biomedical research by using their knowledge and intelligence, performing experiments by hands and eyes, and processing data by basic statistical and mathematical tools. Due to huge amounts of biological data and a very large number of possible combinations and permutations of various biological sequences, the conventional human intelligence-based methods cannot work effectively and efficiently. So artificial intelligence techniques such as machine learning can play a critical role in complex biomedical applications.

Experts from different domains have contributed chapters to this book, which feature novel machine learning methods and their applications in bioinformatics. Relevant machine learning methods include support vector machines, kernel machines, feature selection, neural networks, evolutionary computation, statistical learning, fuzzy logic, supervised learning, clustering, ensemble learning, Bayesian networks, linear regression, principal components analysis, hidden Markov models, entropy-based information methods, and many others. The 20 chapters of the book are organized in a convenient order, based on their contents, so as to enable the readers to easily gather information in a progressive manner. A concise summary of each chapter follows.

In Chapter 1, Kung and Mak present feature selection methods such as the support vector machine recursive feature elimination (SVM-RFE), filter methods, and wrapper methods, in application to microarray data. Filter methods are based on input and output correlation statistics between input and predictions, or signal-to-noise (SNR) statistics, independent of the classifier or predictor. The development of microarray technology has brought with it problems that are interesting, both from statistical and biological perspectives. One important problem is to identify important genes that are relevant to distinguish cancerous samples from benign samples, or different cancer types. In the SVM-RFE, the magnitude of the weight connected to a particular feature is used as the ranking criteria for selection. The methods are illustrated in selection of important genes and in prediction of protein subcellular localization. In the protein subcellular localization, whether a protein lies in the cytoplasm, nuclear, extracellular, mitochondrial, or nuclear location is predicted from its amino acid sequence.
In Chapter 2, Menjoge and Welsch give a new feature selection method using 1-norm SVM and 2-norm SVM techniques, where the weights are used as regularization terms of 1-norm and 2-norm forms. Results show that these methods perform well as compared with other methods. The elastic net, in particular, demonstrates excellent classification accuracy. However, none of the methods dominate the other methods in both selecting a small number of variables and classifying data sets.

In Chapter 3, Kim and Park discuss adaptive supervised machine learning algorithms since the adaptive classifiers avoid expensive recomputation of the solution from scratch. Both an adaptive KDA/RMSE (aKDA/RMSE) based on updating the QR decomposition and an adaptive KDA/MSE based on updating the UTV decomposition KDA/MSE-UTV is proposed. These new kernel classifiers can be applied to compute leave-one-out cross-validation efficiently for bioinformatics applications.

In Chapter 4, Pang, Havukkala, Hu, and Kasabov propose a new gene selection method with better bootstrapping consistency for reliable microarray data analysis. The method ensures the reliability and generalizability of microarray data analysis, which thereby leads to an improvement of disease classification performance. Compared with the traditional gene selection methods without using consistency measurement, bootstrapping consistency method provides more accurate classification results. More importantly, results demonstrate that gene selection with the consistency measurement is able to enhance the reproducibility and consistency in microarray data analysis and proteomics-based diagnostics systems.

In Chapter 5, Wang and Palade introduce a series of fuzzy-based techniques, including the fuzzy gene selection method, the fuzzy C-mean clustering-based enhanced gene selection method, and the neuro-fuzzy ensemble approach for building a microarray cancer classification system. Three benchmark microarray cancer data sets, namely, the leukemia cancer data set, colon cancer data set, and lymphoma cancer data set, are used for simulations. The experimental results show that fuzzy-based systems can be efficient tools for microarray data analysis.

In Chapter 6, Li and Yang provide an ensemble learning method with feature selection to improve generalization performance of single classifiers from three aspects. Experiments on benchmark data show that genetic algorithm-based multitask learning (GA-MTL) is more effective than the earlier heuristic algorithms. The algorithms are demonstrated on a brain glioma data set to show the use of the algorithm as an alternative tool for bioinformatics applications.

In Chapter 7, Ahmad, Singh, Araúzo-Bravo, and Sarai study machine learning methods such as neural networks and support vector machines to predict one-dimensional features of protein structures, such as secondary structure, solvent accessibility, and coordination number, and more recently one-dimensional functional properties such as binding sites. The prediction techniques have been shown to have good performance even in the absence of known homology to other proteins. The computational similarities of the methods are highlighted. Common standards for making such sequence-based predictions are also developed.

In Chapter 8, Bu, Li, Gao, Yu, Xu, and Li give a new protein structure prediction method. Despite significant progresses made recently, every protein structure prediction method still possesses limitations. To overcome such shortcomings, a natural idea
is integrating the strengths of different methods to obtain more accurate structures by boosting some weaker predictors into a stronger one. As suggested by recent CASP competitions, the consensus-based prediction strategies usually outperform others by generating better results.

In Chapter 9, Gubbi, Shilton, and Palaniswami investigate different kernel machines in relation to protein structure prediction. Amino acids arrange themselves in 3D space in stable thermodynamic conformations, referred to as native conformation, and the protein becomes active in this state. Thermodynamic interactions include formation of hydrogen bonding, hydrophobic interactions, electrostatic interactions, and complex formation between metal ions. Protein molecules are quite complex in nature and often made up of repetitive subunits.

In Chapter 10, Jin and Zhang give a new method to predict protein subcellular locations based on SVM with evolutionary granular kernel trees (EGKT) and the one-versus-one voting approach. The new method can effectively incorporate amino acid composition information and combine binary SVM models for protein subcellular location prediction.

In Chapter 11, Liao discusses three applications, where the long-range correlations are believed to be essential, by using specific classification and prediction schemes: hidden Markov models for transmembrane protein topology, stochastic context-free grammars for RNA folding, and global structural profiling for antisense oligonucleotide efficacy. By first examining the limitations of present models, some expansions to capture and incorporate long-range features from the aspects of model architecture, learning algorithms, hybrid models, and model equivalence are made. The performance has been improved consequently.

In Chapter 12, Reddy, Weng, and Chiang give a novel optimization framework that searches the neighborhood regions of the initial alignment in a systematic manner to explore the multiple local optimal solutions. This effective search is achieved by transforming the original optimization problem into its corresponding dynamical system and estimating the practical stability boundary of the local maximum. Results show that the popularly used EM algorithm often converges to suboptimal solutions, which can be significantly improved by the proposed neighborhood profile search.

In Chapter 13, Rajapakse and Ho give a novel approach to encode inputs to neural networks for the recognition of transcription start sites in RNA polymerase II promoter regions. The Markovian parameters are used as inputs to three neural networks, which learn potential distant relationships between the nucleotides at promoter regions. Such an approach allows for incorporating biological contextual information at the promoter sites into neural networks and in general implementing higher-order Markov models of the promoters. Experiments on a human promoter data set show an increased correlation coefficient rate of 0.69 on average, which is better than the earlier reported by the NNPP 2.1 method.

In Chapter 14, Xie, Wu, and Yan propose three eukaryotic promoter prediction algorithms, PromoterExplorer I, II, and III. PromoterExplorer I is developed based on relative entropy and information content. PromoterExplorer II takes different kinds of features as the input and adopts a cascade AdaBoost-based learning procedure to select features and perform classification. The outputs of these two methods are combined to