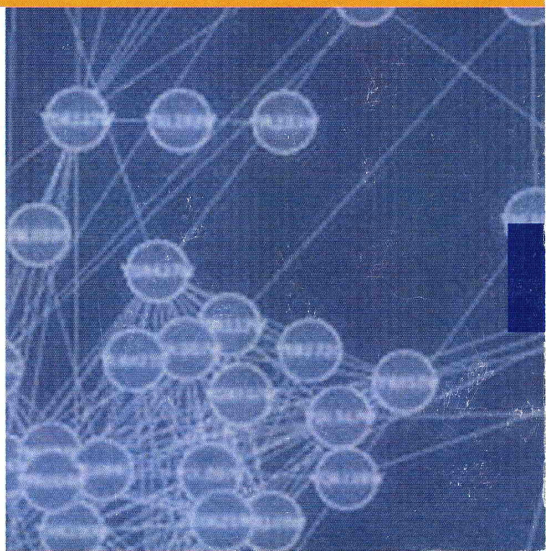
Sašo Džeroski, Pierre Geurts and Juho Rousu (Eds.)

Machine Learning in Systems Biology

Proceedings of The Third International Workshop
Ljubljana, Slovenia
September 5-6, 2009



Sašo Džeroski, Pierre Geurts and Juho Rousu (Eds.)

Machine Learning in Systems Biology

Proceedings of The Third International Workshop
September 5-6, 2009
Ljubljana, Slovenia

Contact Information

Postal address:

Department of Computer Science
P.O.Box 68 (Gustaf Hällströminkatu 2b)
FIN-00014 University of Helsinki
Finland

URL: http://www.cs.helsinki.fi

Telephone: +358 9 1911
Telefax: +358 9 191 51120

Series of Publications B, Report B-2009-1
ISSN 1458-4786
ISBN 978-952-10-5699-4
Computing Reviews (1998) Classification: F.2,I.2.6,G.3,J.3
Helsinki 2009
Helsinki University Printing House

Preface

Molecular biology and all the biomedical sciences are undergoing a true revolution as a result of the emergence and growing impact of a series of new disciplines and tools sharing the '-omics' suffix in their name. These include in particular genomics, transcriptomics, proteomics and metabolomics, devoted respectively to the examination of the entire systems of genes, transcripts, proteins and metabolites present in a given cell or tissue type. The availability of these new, highly effective tools for biological exploration is dramatically changing the way one performs research in at least two respects. First, the amount of available experimental data is not a limiting factor any more; on the contrary, there is a plethora of it. Given the research question, the challenge has shifted towards identifying the relevant pieces of information and making sense out of it (a 'data mining' issue). Second, rather than focus on components in isolation, we can now try to understand how biological systems behave as a result of the integration and interaction between the individual components that one can now monitor simultaneously, so called 'systems biology'.

Machine learning naturally appears as one of the main drivers of progress in this context, where most of the targets of interest deal with complex structured objects: sequences, 2D and 3D structures or interaction networks. At the same time bioinformatics and systems biology have already induced significant new developments of general interest in machine learning, for example in the context of learning with structured data, graph inference, semi- supervised learning, system identification, and novel combinations of optimization and learning algorithms.

This book contains the scientific contributions presented at the Third International Workshop on Machine Learning in Systems Biology (MLSB'2009), held in Ljubljana, Slovenia from September 5 to 6, 2009. The workshop was organized as a core event of the PASCAL2 Network of Excellence, under the IST programme of European Union. The aim of the workshop was to contribute to the cross-fertilization between the research in machine learning methods and their applications to systems biology (i.e., complex biological and medical questions) by bringing together method developers and experimentalists.

The technical program of the workshop consisted of invited lectures, oral presentations and poster presentations. Invited lectures were given by Diego di Bernardo, Roman Jerala, Nick Juty, Yannis Kalaidzidis, Ross D. King, and William Stafford Noble. Twelve oral presentations were given, for which extended abstracts (papers) are included in this book: these were selected from 18 submissions, each reviewed by three members of the scientific program committee. Twenty-two poster presentations were given, for which one-page abstracts are included here. We would like to thank all the people contributing to the technical programme, the scientific program committee, the local organizers and the sponsors for making the workshop possible.

Ljubljana, September 2009 Sašo Džeroski, Pierre Geurts and Juho Rousu

Program Chairs

Sašo Džeroski (Jožef Stefan Institute, Slovenia)
Pierre Geurts (University of Liège, Belgium)
Juho Rousu (University of Helsinki, Finland)

Scientific Program Committee

Florence d'Alché-Buc (University of Evry, France)
Sašo Džeroski (Jožef Stefan Institute, Slovenia)
Paolo Frasconi (Università degli Studi di Firenze, Italy)
Cesare Furlanello (Fondazione Bruno Kessler, Trento, Italy)
Pierre Geurts (University of Liège, Belgium)
Mark Girolami (University of Glasgow, UK)
Dirk Husmeier (Biomathematics & Statistics Scotland, UK)
Samuel Kaski (Helsinki University of Technology, Finland)
Ross D. King (Aberystwyth University, UK)
Neil Lawrence (University of Manchester, UK)
Elena Marchiori (Vrije Universiteit Amsterdam, The Netherlands)
Yves Moreau (Katholieke Universiteit Leuven, Belgium)
William Noble (University of Washington, USA)
Gunnar Rätsch (FML, Max Planck Society, Tübingen)
Juho Rousu (University of Helsinki, Finland)
Céline Rouveirol (University of Paris XIII, France)
Yvan Saeys (University of Gent, Belgium)
Guido Sanguinetti (University of Sheffield, UK)
Ljupčo Todorovski (University of Ljubljana, Slovenia)
Koji Tsuda (Max Planck Institute, Tuebingen)
Jean-Philippe Vert (Ecole des Mines, France)
Louis Wehenkel (University of Liège, Belgium)
Jean-Daniel Zucker (University of Paris XIII, France)
Blaž Zupan (University of Ljubljana, Slovenia)

Local Organizers

Ivica Slavkov (Jožef Stefan Institute, Slovenia)
Dragi Kocev (Jožef Stefan Institute, Slovenia)
Tina Anžič (Jožef Stefan Institute, Slovenia)

Sponsors

PASCAL2 Network of Excellence, Core Event;
IST programme of the European Community, contract IST-2007-216886.
Slovenian Research Agency
Jožef Stefan Institute, Slovenia
University of Helsinki, Finland

Table of Contents

I Invited Lectures

Networking Genes and Drugs: Understanding Drug Mode of Action and Gene Function from Large-scale Experimental Data 1
Diego di Bernardo

Synthetic Biology: Achievements and Prospects for the Future 3
Roman Jerala

Ontologies for Systems Biology 5
Nick Juty

Quantitative Microscopy: Bridge Between "Wet" Biology and Computer Science 7
Yannis Kalaidzidis

On the Automation of Science 9
Ross D. King

Machine Learning Methods for Protein Analyses 11
William Stafford Noble

II Papers

A comparison of AUC estimators in small-sample studies 15
Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets and Tapio Salakoski

Hierarchical cost-sensitive algorithms for genome-wide gene function prediction 25
Nicolo' Cesa-Bianchi and Giorgio Valentini

Evaluation of methods in GA studies: yet another case for Bayesian networks 35
Gábor Hullám, Peter Antal, Csaba Szalai and András Falus

Evaluation of Signaling Cascades Based on the Weights from Microarray and ChIP-seq Data 45
Zerrin Isik, Volkan Atalay and Rengül Çetin-Atala

Matching models to data in modelling morphogen diffusion 55
Wei Liu and Mahesan Niranjan

On utility of gene set signatures in gene expression-based cancer class prediction 65
Minca Mramor, Marko Toplak, Gregor Leban, Tomaž Curk, Janez Demšar and Blaž Zupan

Accuracy-Rejection Curves (ARCs) for Comparison of Classification Methods with Reject Option 75
Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker and Blaise Hanczar

Predicting the functions of proteins in PPI networks from global information 85
Hossein Rahmani, Hendrik Blockeel and Andreas Bender

Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction 95
Matteo Re and Giorgio Valentini

Integrated network construction using event based text mining 105
Yvan Saeys, Sofie Van Landeghem and Yves Van de Peer

Evaluation Method for Feature Rankings and their Aggregations for Biomarker Discovery 115
Ivica Slavkov, Bernard Ženko, and Sašo Džeroski

A Subgroup Discovery Approach for Relating Chemical Structure and Phenotype Data in Chemical Genomics 125
Lan Umek, Petra Kaferle, Mojca Mattiazzi, Aleš Erjavec, Črtomir Gorup, Tomaž Curk, Uroš Petrovič and Blaž Zupan

III Poster Abstracts

Robust biomarker identification for cancer diagnosis using ensemble feature selection methods 135
Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys

Java-ML a Java Library for Data Mining 137
Thomas Abeel, Yves Van de Peer, and Yvan Saeys

Extending KEGG Pathways for a Better Understanding of Prostate Cancer Using Graphical Models 139
Adel Aloraini, James Cussens, and Richard Birnie

Variable Pruning in Bayesian Sequential Study Design 141
P. Antal, G. Hajós, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus

On the Bayesian applicability of graphical models in genome-wide association studies 143
P. Antal, A. Millinghoffer, Cs. Szalai, and A. Falus

Averaging over measurement and haplotype uncertainty using probabilistic genotype data 145
P. Antal, P. Sárközy, B. Zoltán, P. Kiszel, Á. Semsei, Cs. Szalai, and A. Falus

Bayes Meets Boole: Bayesian Learning of Boolean Regulatory Networks from Expression Data 147
Matthias Böck, Soichi Ogishima, Lars Kaderali, and Stefan Kramer

Statistical relational learning for supervised gene regulatory network inference 149
Céline Brouard, Julie Dubois, Marie-Anne Debily, Christel Vrain, and Florence d'Alché-Buc

Top-down phylogenetic tree reconstruction: a decision tree approach 151
Eduardo Costa, Celine Vens, and Hendrik Blockeel

Using biological data to benchmark microarray analysis methods 153
Bertrand De Meulder, Benoît De Hertogh, Fabrice Berger, Anthoula Gaigneaux, Michael Pierre, Eric Bareke,and Eric Depiereux

Structural Modeling of Transcriptomics Data Using Creative Knowledge Discovery 155
Kristina Gruden, Petra Kralj Novak, Igor Mozetič, Vid Podpečan, Matjaž Hren, Helena Motaln, Marko Petek, and Nada Lavrač

Phenotype Prediction from Genotype Data 157
Giorgio Guzzetta, Giuseppe Jurman, Cesare Furlanello

Biomarker Selection by Transfer Learning with Linear Regularized Models 159
Thibault Helleputte, and Pierre Dupont

Combining Semantic Relations from the Literature and DNA Microarray Data for Novel Hypotheses Generation 161
Dimitar Hristovski, Andrej Kastrin, Borut Peterlin, and Thomas C. Rindflesch

Two-Way Analysis of High-Dimensional Metabolomic Datasets 163
Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski

The Open and Closed-World Assumptions in Representing Systems Biology Knowledge 165
Agnieszka Lawrynowicz, Ross D. King

Learning gene networks with sparse inducing estimators 167
Fabian Ojeda, Marco Signoretto, and Johan Suykens

Taking Advantage of the Amount of Archived Affymetrix GeneChips to Identify Genes Involved in Metastasis and Regulated by Hypoxia 169
Michael Pierre, Anthoula Gaigneaux, Bertrand DeMeulder, Fabrice Berger, Benoît DeHertogh, Eric Bareke, Carine Michiels, and Eric Depiereux

Metabolic syndrome assessment using Fuzzy Artmap neural network and 1H NMR spectroscopy 171
Bogdan Pogorelc, Jesus Brezmes, Matjaž Gams

Modeling phagocytosis - PHAGOSYS project outline 173
Barbara Szomolay

Inductive Process-Based Modeling of Endocytosis from Time-Series Data 175
Ljupčo Todorovski and Sašo Džeroski

Analyzing time series gene expression data with predictive clustering rules 177
Bernard Ženko, Jan Struyf, and Sašo Džeroski

Author Index 179

Part I

Invited Lectures

Networking Genes and Drugs: Understanding Drug Mode of Action and Gene Function from Large-scale Experimental Data

Diego di Bernardo[1,2]

[1] Telethon Institute of Genetics and Medicine (TIGEM), Naples 80131, Italy
[2] Department of Computer and Systems Engineering,
University of Naples "Federico II", Naples 80125, Italy

Abstract. A cell can be described as a synergistic ensemble of biological entities (mRNA, proteins, ncRNA, metabolites, etc) interacting with each other, whose collective behaviour causes the observed phenotypes. A great research effort is ongoing in identifying and mapping the network of interactions among biomolecules in mammalian species. The idea of harnessing this network to understand human diseases at the molecular level, and possibly to find suitable drugs for their treatment, is fascinating but still unfulfilled. We will show how it is possible to harness experimental data on human cells and tissue to identify the gene regulatory networks among tens of thousands of genes, and how to use this information to analyse the modular structure of the cell and predict the function of each gene. Moreover, we will show how using these data it is also possible to identify a suitable drug, or a combination of drugs, that can restore the physiological behaviour of the affected pathways in human diseases.

Synthetic Biology: Achievements and Prospects for the Future

Roman Jerala[1,2]

[1] Department of Biotechnology, National institute of Chemistry, Ljubljana, Slovenia
[2] Faculty of Chemistry and chemical technology, University of Ljubljana, Slovenia

Abstract. Synthetic biology, which combines engineering approach in biological systems is getting a strong momentum due to the recent technological advances, which allow us to manipulate the genetic information at an unprecedented scale. Currently synthetic biology is exploiting its potentials and advantages but also bottlenecks. We will review some success stories of synthetic biology in different field of applications, such as medicine, energy and materials. Medical applications of synthetic biology are some of the most promising areas of synthetic biology, particularly for the alternative methods of drug production, biosensors and also different therapeutic applications. Recent developments in our understanding of cellular signaling and host-pathogen interactions provide the opportunity for new types of medical intervention, where we can utilize parts of the existing or re-engineer signaling responses connected to various pathological conditions. Knowledge of the ways that microbes use to avoid the human immune response allows us to devise approach to bypass those microbial strategies. We will look at three different applications of synthetic biology, which involve re-engineering of cell signaling pathways, which we have prepared for the international genetically engineered machines competition in years 2006-2008. We have designed and demonstrated proof of the concept of antiviral detection and defense system based on essential viral functions that is independent on mutations and a synthetic vaccine that activates both innate and adaptive immune response.

Ontologies for Systems Biology

Nick Juty

EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, United Kingdom

Abstract. The ease with which modern computational and theoretical tools can be applied to modeling has led to an exponential increase in the size and complexity of computational models in biology. At the same time, the accelerating pace of progress also highlights limitations in current approaches to modeling. One of these limitations is the insufficient degree to which the semantics and qualitative behaviour of models are systematised and expressed formally enough to support unambiguous interpretation by software systems. As a result, human intervention is required to interpret and connect a model's mathematical structures with information about the its meaning (semantics). Often, this critical information is usually communicated through free-text descriptions or non-standard annotations; however, free-text descriptions cannot easily be interpreted by current modeling tools.

We will describe three efforts to standardise the encoding of missing semantics for kinetic models. The overall approach involves connecting model elements to common, external sources of information that can be extended as existing knowledge is expanded and refined. These external sources are carefully managed public, free, consensus ontologies: the Systems Biology Ontology (SBO), the Kinetic Simulation Algorithm Ontology (KiSAO), and the Terminology for the Description of Dynamics (TeDDy). Together they provide a means for annotating a model with stable and perennial identifiers which reference machine-readable regulated terms defining the semantics of the three facets of the modeling process: 1. the relationship between the model and the biology it aims to describe, 2. the process used to simulate the model and obtain expected results, and 3. the results themselves.

Quantitative Microscopy: Bridge Between "Wet" Biology and Computer Science

Yannis Kalaidzidis

Max Plank Institute of Molecular Cell Biology and Genetics
Pfotenhauerstrasse 108, 01307, Dresden, Germany

Abstract. Quantification of experimental evidence is an important aspect of modern life science. In microscopy, this causes a shift from pure presentation of "supporting cases" toward the quantification of the processes under study. Computer image processing breaks through the light microscopy diffraction limit, it allows to track individual molecules in the life specimen, quantify distribution and co-localization of compartment markers, etc. The quantified experimental data forms a basis for the models of the biological processes. Quality of predictive models is crucially dependent on the accuracy of the quantified experimental data. The quality of experimental data is a function of algorithms as well as the imperfections of the "wet" experiment. The number of research papers devoted to the algorithms of microscopy image analysis, segmentation, classification and tracking has grown very fast in the last two decades. The analysis of the source of noise in "wet" biology and microscopy has gotten less attention. In this talk I will focus on the correction of experimental data before applying analysis algorithms. These corrections have two faces. They are obligatory to compensate for imperfections of "wet" microscopy while at the same time this correction can break some assumptions, which form the basis of algorithms for subsequent analysis. The examples of the different approaches for "pre-" and "post-" correction will be presented.

On the Automation of Science

Ross D. King

Department of Computer Science, Aberystwyth University, Wales, UK

Abstract. The basis of science is the hypothetico-deductive method and the recording of experiments in sufficient detail to enable reproducibility. We report the development of the Robot Scientist "Adam" which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast Saccharomyces cerevisiae, and experimentally tested these hypotheses using laboratory automation. We have confirmed Adam's conclusions through manual experiments. To describe Adam's research we have developed an ontology and logical language. The resulting formalization involves over 10,000 different research units in a nested tree-like structure, ten levels deep, that relates the 6.6 million biomass measurements to their logical description. This formalization describes how a machine discovered new scientific knowledge. Describing scientific investigations in this way opens up new opportunities to apply machine learning and data-mining to discover new knowledge.

Machine Learning Methods for Protein Analyses

William Stafford Noble[1,2]

[1] Department of Genome Sciences, University of Washington, Seattle, WA 98195
[2] Department of Computer Science and Engineering
University of Washington, Seattle, WA 98195

Abstract. Computational biologists, and biologists more generally, spend a lot of time trying to more fully characterize proteins. In this talk, I will describe several of our recent efforts to use machine learning methods to gain a better understanding of proteins. First, we tackle one of the oldest problems in computational biology, the recognition of distant evolutionary relationships among protein sequences. We show that by exploiting a global protein similarity network, coupled with a latent space embedding, we can detect remote protein homologs more accurately than state-of-the-art methods such as PSI-BLAST and HHPred. Second, we use machine learning methods to improve our ability to identify proteins in complex biological samples on the basis of shotgun proteomics data. I will describe two quite different approaches to this problem, one generative and one discriminative.

Part II

Papers

A comparison of AUC estimators in small-sample studies

Antti Airola,[1] Tapio Pahikkala,[1] Willem Waegeman,[2] Bernard De Baets,[2] and Tapio Salakoski[1]

[1] Department of Information Technology, University of Turku and Turku Centre for Computer Science (TUCS), Joukahaisenkatu 3-5 B, Turku, Finland
[2] KERMIT, Department of Applied Mathematics, Biometrics and Process Control, Coupure links 653, Ghent University, Belgium

Abstract. Reliable estimation of the classification performance of learned predictive models is difficult, when working in the small sample setting. When dealing with biological data it is often the case that separate test data cannot be afforded. Cross-validation is in this case a typical strategy for estimating the performance. Recent results, further supported by experimental evidence presented in this article, show that many standard approaches to cross-validation suffer from extensive bias or variance when the area under ROC curve (AUC) is used as performance measure. We advocate the use of leave-pair-out cross-validation (LPOCV) for performance estimation, as it avoids many of these problems.

1 Introduction

Small-sample biological datasets, such as microarray data, exhibit properties which pose serious challenges for reliable evaluation of the quality of prediction functions learned from this data. It is typical for genomic studies to produce data containing thousands of features, measured from a small sample of possibly only tens of examples. Further, the relative distribution of the classes to be predicted is often highly imbalanced and their discriminability can be quite low.

AUC is a ranking-based measure of classification performance, which has gained substantial popularity in the machine learning community during recent years [1–3]. Its value can be interpreted as the probability that a classifier is able to distinguish a randomly chosen positive example from a randomly chosen negative example. In contrast to many alternative performance measures, AUC is invariant to relative class distributions, and class-specific error costs. These properties have prompted the use of the AUC measure in microarray studies [4, 5], medical decision making [6], and evaluation of biomedical text mining systems [7] to name a few examples.

When setting aside data for parameter estimation and validation of results cannot be afforded, cross-validation is typically used. However, in [8] it was shown that when considering AUC in the small-sample setting, many commonly used cross-validation schemes suffer from substantial negative bias. In this work, we

explore this issue further and propose LPOCV, first considered in [9] for ranking tasks, as an approach that provides an almost unbiased estimate of expected AUC performance, and also does not suffer from as high variance as some of the alternative strategies.

2 Performance Estimation

Let D be a probability distribution over a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where the input space $\mathcal{X}$ is a set and the output space $\mathcal{Y} = \{-1, 1\}$. An example $z = (x, y) \in \mathcal{Z}$ is thus a pair consisting of an input and an associated label, which describes whether the example belongs to the positive or to the negative class. The conditional distribution of an input from $\mathcal{X}$, given that it belongs to the positive class is denoted by D_+, and given that it belongs to the negative class by D_-. Further, let the sequence $Z = ((x_1, y_1), \ldots, (x_m, y_m)) \in \mathcal{Z}^m$ drawn independent and identically distributed from D be a training set of m training examples, with $X = (x_1, \ldots, x_m) \in \mathcal{X}^m$ denoting the inputs and $Y = (y_1, \ldots, y_m) \in \mathcal{Y}^m$ the labels in the training set.

Now let us consider a prediction function f_Z returned by a learning algorithm based on a fixed training set Z. We are interested in the generalization performance of this function, that is, how well it will predict on unseen future data. The generalization performance of f_Z can be measured by its expected AUC $A(f_Z)$, sometimes also known as expected ranking accuracy [10], over all possible positive-negative example pairs, that is

$$A(f_Z) = E_{x_+ \sim D_+ x_- \sim D_-}[H(f_Z(x_+) - f_Z(x_-))]$$

where H is the Heaviside step function, for which $H(a)$ is 1 if $a > 0$, 1/2 if $a = 0$, and 0 if $a < 0$. We call this measure the *conditional expected AUC* of the prediction function, as it is conditioned on a fixed training set Z.

Alternatively, we may also want to consider the expectation taken over all possible training sets of size m. The *unconditional expected AUC* can be defined as

$$E_{Z \sim D^m}[A(f_Z)].$$

As discussed for example in [11, 12], these two measures correspond to two different questions of interest. The conditional expected performance corresponds to the question how well we expect that a prediction function learned from a given training set will generalize to future examples. The unconditional expected performance measures the quality of the learning algorithm itself, that is, how well on average will a prediction function learned by the algorithm of interest from a dataset of a given size generalize to new data.

More often, machine learning related articles concentrate on the unconditional performance, as the goal usually is to measure the quality of learning algorithms, where the training data is treated as a random variable. However, as argued by [11], the conditional error estimate is more of interest in a setting

where a researcher is using a certain dataset and wants to know how well a prediction function learned from that particular dataset will do on future examples. This is the setting we concentrate on in this paper.

In practice we almost never can directly access the probability distribution D to calculate A, but are rather limited to using some estimate $\hat{A}$ instead. To measure the quality of an estimator, in terms of its ability to measure conditional expected AUC, we follow the setting of [11]. We consider the deviation $B(Z) = \hat{A}(f_Z) - A(f_Z)$, which measures the difference between the estimated and true conditional expected AUC of a prediction function.

We study the expected value $E_{Z\sim D^m}[B(Z)]$ of the deviation distribution as a measure of the biasedness of the estimator. Further, we consider the variance $\mathrm{Var}_{Z\sim D^m}[B(Z)]$ of the deviation distribution, as a measure of the reliability of individual estimates. Preferably an estimator would have both close to zero deviation mean and variance.

The AUC measure can be calculated using the following formula, also called the Wilcoxon-Mann-Whitney statistic:

$$\hat{A}(S, f_Z) = \frac{1}{|S_+||S_-|} \sum_{x_i \in S_+} \sum_{x_j \in S_-} H(f_Z(x_i) - f_Z(x_j)),$$

where S is a sequence of examples, and $S_+ \subset S$ and $S_- \subset S$ denote the positive and negative examples in S, respectively. (for proof, see [13]).

In this paper, we consider a commonly used performance evaluation technique known as cross-validation. Here, the dataset is repeatedly partitioned into two non-overlapping parts, a training set and a hold-out set. For each partitioning, the hold-out set is used for testing while the remainder is used for training. The two most popular variants are *tenfold cross-validation*, where the data is split into ten mutually disjoint folds, and *leave-one-out cross-validation* (LOOCV), where each training example constitutes its own fold.

Stratification is commonly done to ensure that the hold-out sets share approximately the same class distributions. Further, for stratified CV on small datasets [8] has recently suggested a balancing strategy to ensure that all the training sets share the same number of positive and negative examples. When the sample size for a class is not a multiple of the number of folds, some folds will contain one extra example from that class compared to the other folds. The balancing is done by randomly removing members of overrepresented classes on each round of cross-validation, so that all the training sets contain the same number of examples from each class.

As discussed in [1, 8], two alternative strategies can be used to calculate the cross-validation estimate over the folds, *pooling* and *averaging*.

In pooling, the predictions made in each cross-validation round are pooled into a one set and one common AUC score is calculated from it. For LOOCV this is the only way to obtain the AUC score. The assumption made when using pooling is that classifiers produced on different cross-validation rounds come from the same population. This assumption may make sense when using performance measures such as classification accuracy, but it is more dubious

when computing AUC, since some of the positive-negative pairs are constructed using data instances from different folds. Indeed, [8] show that this assumption is generally not valid for cross-validation and can lead to large pessimistic biases. In their experiments with no-signal data sets, AUC values of less than 0.3 were observed instead of the expected 0.5.

An alternative approach, averaging, is to calculate the AUC score separately for each cross-validation fold and average them to obtain one common performance estimate. However, the number of positive-negative example pairs in the folds may be too small for calculating AUC reliably when using small imbalanced datasets. As an extreme case, if there are more folds than observations for the minority class, then some of the folds cannot have examples from this class. For such folds, the AUC cannot be calculated.

LPOCV [9, 14] was first introduced for general ranking tasks. Here, we propose its use for AUC calculation, since it avoids many of the pitfalls associated with the pooling and averaging techniques. Analogously to LOOCV, each possible positive-negative pair of training instances is left out of at a time from the training set. Formally, the AUC performance is calculated with LPOCV as

$$\frac{1}{|X_+||X_-|} \sum_{x_i \in X_+} \sum_{x_j \in X_-} H(f_{\overline{\{i,j\}}}(x_i) - f_{\overline{\{i,j\}}}(x_j)),$$

where $f_{\overline{\{i,j\}}}$ denotes a classifier trained without the i-th and j-th training example. Being an extreme form of averaging, where each positive-negative pair of training examples forms an individual hold-out set, this approach is natural when AUC is used as a performance measure, since it guarantees the maximal use of available training data. Moreover, the LPOCV estimate, taken over a training set of m examples, is an unbiased estimate of the unconditional expected AUC over a sample of $m - 2$ examples (for a proof, see [9]).

The computational cost can be seen as a limitation for cross-validation techniques in general, and in particular for the LOOCV and LPOCV. For a training set of m examples a straightforward implementation of LOOCV requires training the learner m times, with LPOCV the required number of training rounds is of the order $O(m^2)$. While these computational costs may be affordable on small training sets, they can become a limiting factor as the training set size increases.

However, for regularized least-squares (RLS) [15] and the AUC-maximizing ranking RLS (RankRLS) [16], efficient algorithms for cross-validation can be derived using techniques based on matrix calculus [17, 14]. Since these algorithms have state-of-the-art classification performance similar to that of the Support Vector Machine (SVM), and Ranking SVM (see e.g. [18, 16]), they are a natural choice to use in settings where cross-validation is important.

3 Empirical study

In the simulation study, we measure the mean and variance of the deviation distribution of several different cross-validation estimators. We consider three

pooled strategies; LOOCV, balanced LOOCV and pooled tenfold, as well as the averaged fivefold, tenfold and LPOCV. Stratification is used where possible.

Our setting is similar to that of [8], where the bias of pooling and averaging approaches was compared on low-dimensional data. We consider synthetic data, as this allows estimating the conditional expected AUC of the learned prediction functions. The training set size is 30 examples in all the simulations, the relative distribution of positive examples is varied between 10% and 90% on 10 percentage unit intervals. We consider both low-dimensional data with 10 features, and high-dimensional data with 1000 features.

In the no-signal experiment, there is no difference between the two classes. Examples from both classes are sampled from normal distributions with zero mean, unit variance and no covariance between the features. The conditional expected AUC of a prediction function is in this setting 0.5, as no model can do either better or worse than random, in terms of AUC. In the signal experiment the means of a number of features are shifted to 0.5 for the positive, and to -0.5 for the negative class. With 10 features, 1 feature is shifted, with 1000 features, 10 features are shifted. Generated test sets with 10000 examples are used to estimate the conditional expected AUC of the learned prediction functions.

Two learning algorithms are considered in the experiments, RLS and RankRLS. RLS optimizes an approximation of accuracy, like most machine learning algorithms, while RankRLS optimizes more directly the AUC. We only investigated the linear kernel, since in bioinformatics it is commonly assumed that high-dimensional data can be separated in a linear way. The considered learners have also a regularization parameter, which controls the tradeoff between model complexity and fit to the training data. In the experiments we did not find the level of regularization applied to have major effect on the relative quality of the cross-validation estimates, so we consider only the results for regularization parameter value 1. The used learning and cross-validation algorithms are from our RLScore software package, available at `http://www.tucs.fi/rlscore`. All the experiments are repeated 10000 times. We assess the significance of the difference between the deviation of the LPOCV estimate and the alternative estimates using the Wilcoxon signed-rank test, with $p = 0.05$, applying the Bonferroni correction for multiple hypothesis testing.

Figure 1 displays the results for non-signal data. When using the RLS- learner on low-dimensional data, we observe a substantial bias for the pooled estimators, with balanced LOOCV being the least biased of them. The averaging strategies work better, with LPOCV showing significantly less bias than all of the pooled strategies. These results are consistent with those reported in [8]. With RankRLS and low-dimensional data, the pessimistic bias of the pooled strategies is much smaller, but nonetheless significant differences compared to the less pessimistic LPOCV are observed. LPOCV and the other averaged strategies behave similarly. On high-dimensional data none of the estimates show clear bias.

Figure 2 displays the results for signal data. Again, with the RLS learner and low-dimensional data, a large pessimistic bias is present in the pooled estimates. LPOCV gives significantly less biased performance estimates. For RankRLS we

observe the same phenomenon, though the negative bias of the pooled strategies is much smaller than for RLS (similarly to the no-signal experiment). On high-dimensional data, most of the pessimistic bias seems to disappear from the pooled estimates. With RankRLS, LOOCV actually provides significantly more optimistic performance estimates than LPOCV, though the magnitudes of the differences in their mean deviations are very small. Of the averaged strategies, the bias of tenfold cross-validation is similar to that of LPOCV. However, averaged fivefold cross-validation is in most of the signal experiments much more pessimistically biased than LPOCV.

In all of the experiments, averaged tenfold and fivefold strategies have larger variance than the pooled strategies and LPOCV. The more imbalanced the relative class distributions, the higher the variance becomes. This effect is magnified for averaged tenfold and fivefold, as folds which do not have examples from both classes can not be considered when calculating the average AUC.

To conclude, LPOCV shows very little bias in both low- and high dimensional feature space, and has a very similar variance to that of the pooled strategies. Averaged tenfold cross-validation is also very competitive in terms of bias, but suffers from large variance, as does averaged fivefold cross-validation. Furthermore, for averaged fivefold large pessimistic bias appears in the signal experiment. This is probably due to the fact that one fifth of the training data is held out of the already very small training set in each round. LOOCV and balanced LOOCV worked well in many settings, but both suffered from a large negative bias on low-dimensional data and RLS learner.

4 Conclusion

In this work we have considered the merits and drawbacks of different conditional expected AUC cross-validation estimators, in the small sample setting. In terms of variance, the averaged fivefold and tenfold cross-validation proved to be inferior to the pooled strategies and LPOCV. On low dimensional data sets, large negative bias was observed in the pooled estimators showing that they can systematically fail in such a setting. However, with increased dimensionality this effect disappeared, suggesting that the pooled estimators can be very competitive when using high dimensional data. LPOCV seems to be overall the most robust method, as it is in all settings almost unbiased, and shows variance that is competitive with that of the pooled estimators.

Based on the simulation results we suggest the use of LPOCV for AUC-estimation due to its robustness. For RLS based learners calculating the LPOCV can be done efficiently, for other types of methods the computational cost can be high. Further study is needed to ascertain whether the large bias exhibited by the pooled estimators is a phenomenom that appears only when dealing with small dimensional data. If this is the case, the pooled CV strategies may also be considered suitable for AUC estimation for high dimensional data, which is a typical property of data produced by biomolecular studies.

Acknowledgments

This work has been supported by the Academy of Finland. W.W. was supported by a research visit grant from the Research Foundation Flanders.

References

1. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. **30**(7) (1997) 1145–1159
2. Waegeman, W., De Baets, B., Boullart, L.: ROC analysis in ordinal regression learning. Pattern Recogn. Lett. **29**(1) (2008) 1–9
3. Vanderlooy, S., Hüllermeier, E.: A critical analysis of variants of the AUC. Mach. Learn. **72**(3) (2008) 247–262
4. Baker, S., Kramer, B.: Identifying genes that contribute most to good classification in microarrays. BMC Bioinformatics **7**(1) (2006)
5. Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., De Moor, B.: Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. Bioinformatics **22**(14) (2006) 184–190
6. Swets, J.: Measuring the accuracy of diagnostic systems. Science **240**(4857) (1988) 1285–1293
7. Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., Salakoski, T.: All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics **9**(Suppl 11) (2008) S2
8. Parker, B.J., Gunter, S., Bedo, J.: Stratification bias in low signal microarray studies. BMC Bioinformatics **8**(326) (2007)
9. Cortes, C., Mohri, M., Rastogi, A.: An alternative ranking problem for search engines. In: Proceedings of WEA'07. (2007) 1–21
10. Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., Roth, D.: Generalization bounds for the area under the ROC curve. J. Mach. Learn. Res. **6** (2005) 393–425
11. Braga-Neto, U.M., Dougherty, E.R.: Is cross-validation valid for small-sample microarray classification? Bioinformatics **20**(3) (2004) 374–380
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, Second Edition. (2009)
13. Cortes, C., Mohri, M.: AUC optimization vs. error rate minimization. In Thrun, S., Saul, L., Schölkopf, B., eds.: Proceedings of NIPS'03. (2003)
14. Pahikkala, T., Airola, A., Boberg, J., Salakoski, T.: Exact and efficient leave-pair-out cross-validation for ranking RLS. In Honkela, T., Pöllä, M., Paukkeri, M.S., Simula, O., eds.: Proceedings of AKRR'08. (2008) 1–8
15. Rifkin, R., Yeo, G., Poggio, T.: Regularized least-squares classification. In Suykens, J., Horvath, G., Basu, S., Micchelli, C., Vandewalle, J., eds.: Advances in Learning Theory: Methods, Model and Applications. (2003) 131–154
16. Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., Järvinen, J.: An efficient algorithm for learning to rank from preference graphs. Mach. Learn. **75**(1) (2009) 129–165
17. Pahikkala, T., Boberg, J., Salakoski, T.: Fast n-fold cross-validation for regularized least-squares. In Honkela, T., Raiko, T., Kortela, J., Valpola, H., eds.: Proceedings of SCAI'06. (2006) 83–90
18. Zhang, P., Peng, J.: SVM vs regularized least squares classification. In Kittler, J., Petrou, M., Nixon, M., eds.: Proceedings of ICPR'04. (2004) 176–179

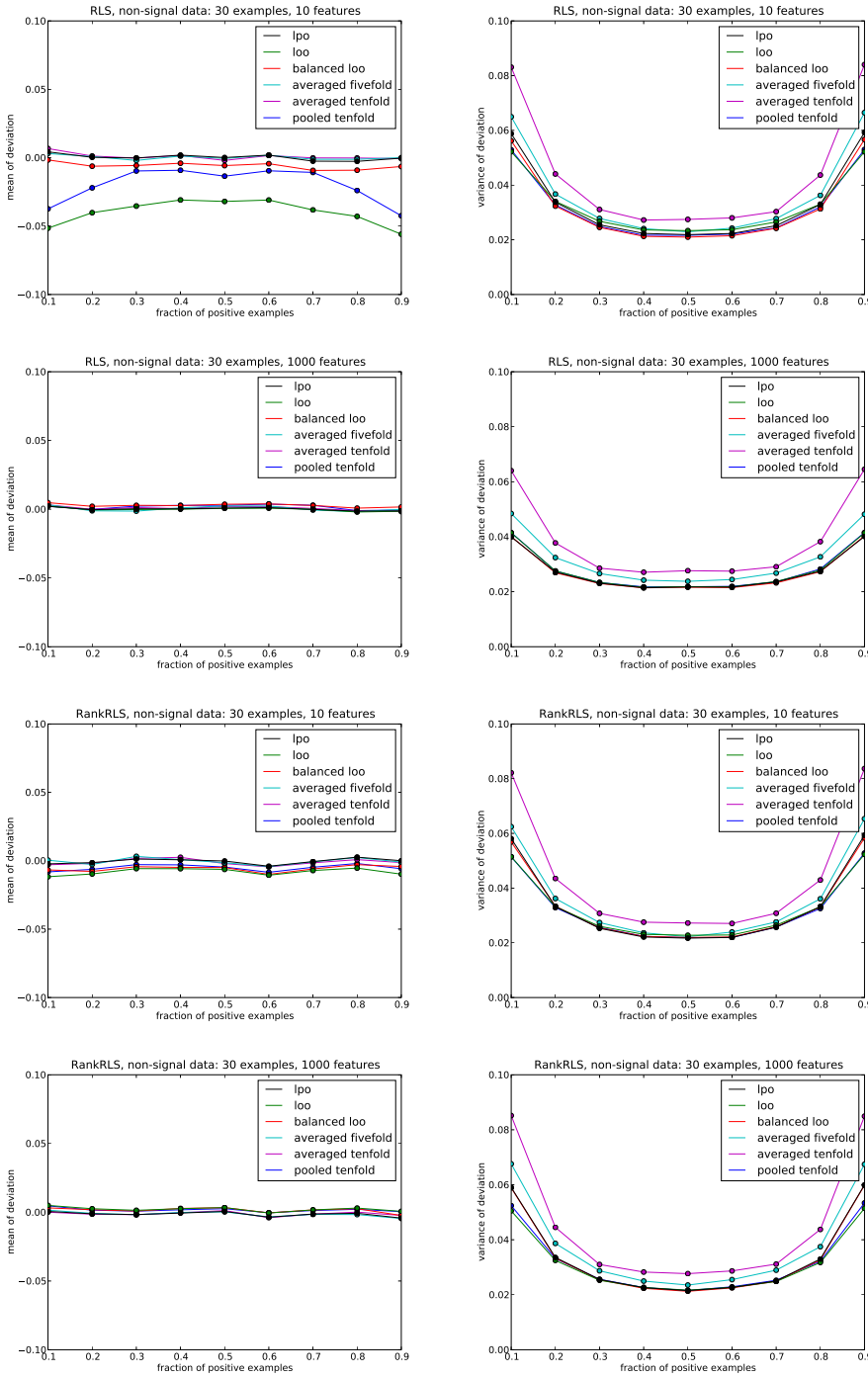


Fig. 1. Mean and variance of the deviation distribution for the non-signal data.

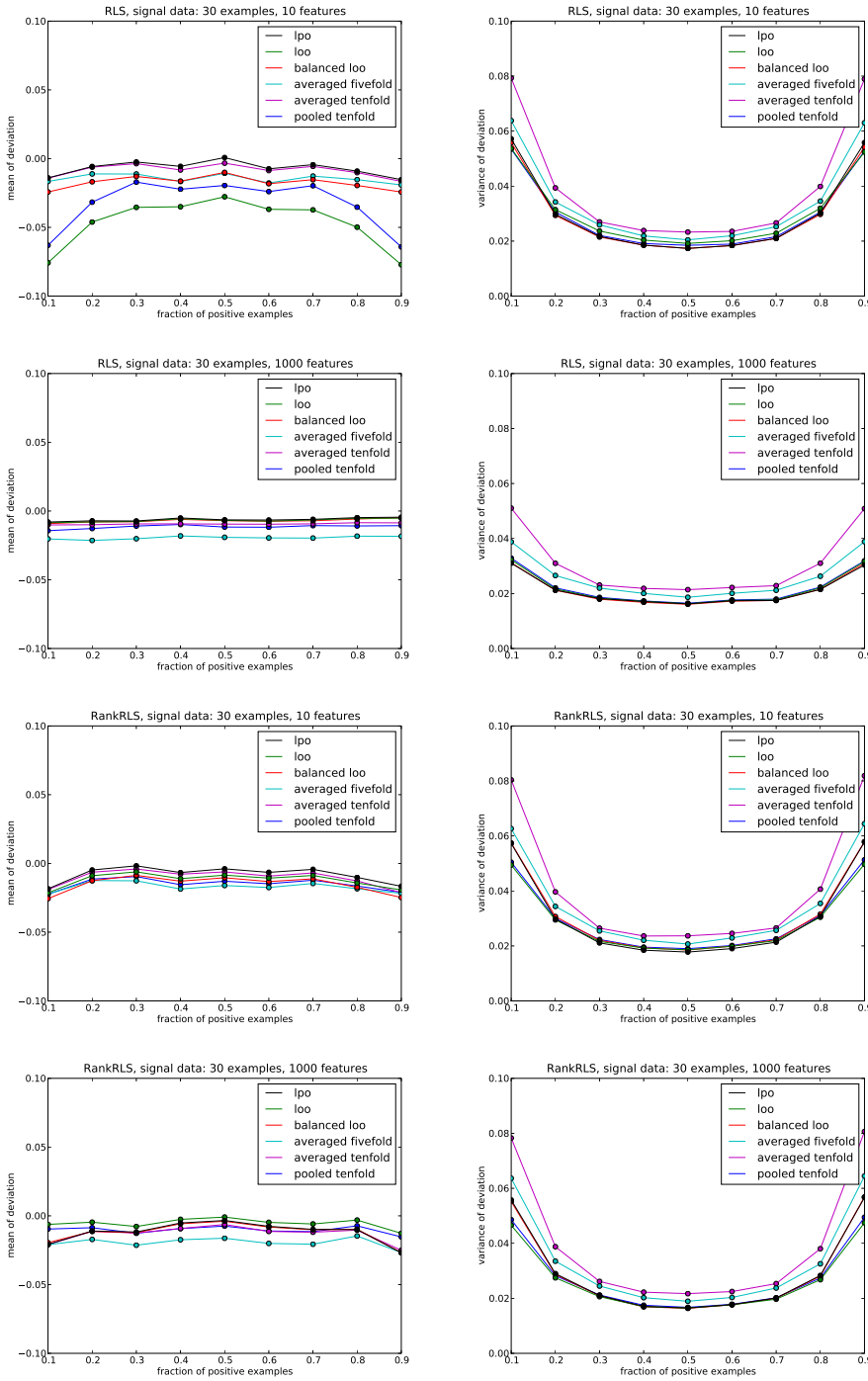


Fig. 2. Mean and variance of the deviation distribution for the signal data.

Hierarchical cost-sensitive algorithms for genome-wide gene function prediction

Nicolò Cesa-Bianchi and Giorgio Valentini

DSI, Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
Via Comelico 39, 20135 Milano, Italia
{cesa-bianchi,valentini}@dsi.unimi.it

Abstract. In this work we propose new ensemble methods for the hierarchical classification of gene functions. Our methods exploit the hierarchical relationships between the classes in different ways: each ensemble node is trained "locally", according to its position in the hierarchy; moreover, in the evaluation phase the set of predicted annotations is built so to minimize a global loss function defined over the hierarchy. We also address the problem of sparsity of annotations by introducing a cost-sensitive parameter that allows to control the precision-recall trade-off. Experiments with the model organism *S. cerevisiae*, using the FunCat taxonomy and 7 biomolecular data sets, reveal a significant advantage of our techniques over "flat" and cost-insensitive hierarchical ensembles.

1 Introduction

"In silico" gene function prediction can generate hypotheses to drive the biological discovery and validation of gene functions. Indeed, "in vitro" methods are costly in time and money, and the computational prediction can support the biologist in understanding the role of a protein or of a biological process, or in annotating a new genome at high level of accuracy, or more in general in solving problems in functional genomics.

Gene function prediction is a classification problem with the following distinctive features: (a) a large number of classes, with multiple functional annotations for each gene (a multiclass multilabel classification problem); (b) hierarchical relationships between classes governed by the "true path rule" [1]; (c) unbalance between positive and negative examples for most classes (sparse multilabels); (d) uncertainty of labels and incompleteness of annotations; (e) availability and need of integration of multiple sources of data.

This paper focuses on the three first items, proposing an ensemble approach for the hierarchical cost-sensitive classification of gene functions at genome and ontology-wide level. Indeed, in this context "flat" methods may introduce large inconsistencies in parent-child relationships between classes, and a hierarchical approach may correct "flat" predictions in order to improve the accuracy and the consistency of the overall annotations of genes [2]. We propose a hierarchical bottom-up Bayesian cost-sensitive ensemble that on the one hand respects

the consistency of the taxonomy, and on the other hand exploits the hierarchical relationships between the classes. Our approach also takes into account the sparsity of annotations in order to improve the precision and the recall of the predictions. We also propose a simple variant of the hierarchical top-down algorithm that optimizes the decision threshold for maximizing the F-score.

Different research lines have been proposed for the hierarchical prediction of gene functions, ranging from structured-output methods, based on the joint kernelization of both input variables and output labels [3, 4], to ensemble methods, where different classifiers are trained to learn each class, and then combined to take into account the hierarchical relationships between functional classes [2, 5]. Our work goes along this latter line of research, and our main contribution is the introduction of a global cost-sensitive approach and the adaptation of a Bayesian bottom-up method to the hierarchical prediction of gene functions using the FunCat taxonomy [6].

Notation and terminology. We identify the N functional classes of the FunCat taxonomy with the nodes $i = 1, \ldots, N$ of a tree T. The root of T is a dummy class with index 0, which every gene belongs to, that we added to facilitate the processing. The FunCat *multilabel* of a gene is the nonempty subset of $\{1, \ldots, N\}$ corresponding to all FunCat classes that can be associated with the gene. We denote this subset using the incidence vector $\boldsymbol{v} = (v_1, \ldots, v_N) \in \{0,1\}^N$. The multilabel of a gene is built starting from the set of terms occurring in the gene's FunCat annotation. As these terms correspond to the most specific classes in T, we add to them all the nodes on paths from these most specific nodes to the root. This "transitive closure" operation ensures that the resulting multilabel satisfies the true path rule. Conversely, we say that a multilabel $\boldsymbol{v} \in \{0,1\}^N$ respects T if and only if $\boldsymbol{v}$ is the union of one or more paths in T, where each path starts from a root but need not terminate on a leaf. All the hierarchical algorithms considered in this paper generate multilabels that respect T. Finally, given a set of d features, we represent a gene with the normalized (unit norm) vector $\boldsymbol{x} \in \mathbb{R}^d$ of its feature values.

2 Methods

The HBAYES ensemble method [7, 8] is a general technique for solving hierarchical classification problems on generic taxonomies. The method consists in training a calibrated classifier at each node of the taxonomy. This is used to derive estimates $\widehat{p}_i(\boldsymbol{x})$ of the probabilities $p_i(\boldsymbol{x}) = \mathbb{P}\big(V_i = 1 \mid V_{\mathrm{par}(i)} = 1, \boldsymbol{x}\big)$ for all $\boldsymbol{x}$ and i, where $(V_1, \ldots, V_N) \in \{0,1\}^N$ is the vector random variable modeling the multilabel of a gene $\boldsymbol{x}$ and $\mathrm{par}(i)$ is the unique parent of node i in T. In order to enforce that only multilabels $\boldsymbol{V}$ that respect T should have nonzero probability, the base learner at node i is only trained on the subset of the training set including all examples $(\boldsymbol{x}, \boldsymbol{v})$ such that $v_{\mathrm{par}(i)} = 1$.

In the evaluation phase, HBAYES predicts the Bayes-optimal multilabel $\widehat{\boldsymbol{y}} \in \{0,1\}^N$ for a gene $\boldsymbol{x}$ based on the estimates $\widehat{p}_i(\boldsymbol{x})$ for $i = 1, \ldots, N$. Namely,

$\widehat{\boldsymbol{y}} = \operatorname{argmin}_{\boldsymbol{y}} \mathbb{E}\big[\ell_H(\boldsymbol{y}, \boldsymbol{V}) \mid \boldsymbol{x}\,\big]$, where the expectation is w.r.t. the distribution of $\boldsymbol{V}$. Here $\ell_H(\boldsymbol{y}, \boldsymbol{V})$ denotes the H-loss [7, 8], measuring a notion of discrepancy between the multilabels $\boldsymbol{y}$ and $\boldsymbol{V}$. The main intuition behind the H-loss is simple: *if a parent class has been predicted wrongly, then errors in its descendants should not be taken into account*. Given fixed *cost coefficients* $c_1, \ldots, c_N > 0$, $\ell_H(\widehat{\boldsymbol{y}}, \boldsymbol{v})$ is computed as follows: all paths in the taxonomy T from the root 0 down to each leaf are examined and, whenever a node $i \in \{1, \ldots, N\}$ is encountered such that $\widehat{y}_i \neq v_i$, then c_i is added to the loss, while all the other loss contributions from the subtree rooted at i are discarded. As shown in [8], $\widehat{\boldsymbol{y}}$ can be computed via a simple bottom-up message-passing procedure whose only parameters are the probabilities $\widehat{p}_i(\boldsymbol{x})$.

We now describe a simple cost-sensitive variant, HBAYES-CS, of HBAYES, which is suitable for learning datasets whose multilabels are sparse. This variant introduces a parameter α that is used to trade-off the cost of false positive (FP) and false negative (FN) mistakes. We start from an equivalent reformulation of the HBAYES prediction rule

$$\widehat{y}_i = \operatorname*{argmin}_{y \in \{0,1\}} \left(c_i^- p_i(1-y) + c_i^+ (1-p_i) y + p_i \{y = 1\} \sum_{j \in \text{child}(i)} H_j \right) \qquad (1)$$

where $H_j = c_j^- p_j (1 - \widehat{y}_j) + c_j^+ (1 - p_j) \widehat{y}_j + \sum_{k \in \text{child}(j)} H_k$ is recursively defined over the nodes j in the subtree rooted at i with each $\widehat{y}_j$ set according to (1), and $\{\, A \,\}$ is the indicator function of event A. Furthermore, $c_i^- = c_i^+ = c_i/2$ are the costs associated to a FN (resp., FP) mistake. In order to vary the relative costs of FP and FN, we now introduce a factor $\alpha \geq 0$ such that $c_i^- = \alpha c_i^+$ while keeping $c_i^+ + c_i^- = 2c_i$. Then (1) can be rewritten as

$$\widehat{y}_i = 1 \iff p_i \left(2c_i - \sum_{j \in \text{child(i)}} H_j \right) \geq \frac{2c_i}{1+\alpha} \ .$$

This is the rule used by HBAYES-CS in our experiments.

Given a set of trained base learners providing estimates $\widehat{p}_1, \ldots, \widehat{p}_N$, we compare the quality of the multilabels computed by HBAYES-CS with that of HTD-CS, a standard top-down hierarchical ensemble method with a cost sensitive parameter $\tau > 0$. The multilabel predicted by HTD-CS is defined by

$$\widehat{y}_i = \{\widehat{p}_i(\boldsymbol{x}) \geq \tau\} \times \{\widehat{y}_{\text{par}(i)} = 1\}$$

for $i = 1, \ldots, N$ (we assume that the guessed label $\widehat{y}_0$ of the root of T is always 1). Note that both methods use the same estimates $\widehat{p}_i$. The only difference is in the way the classifiers are defined in terms of these estimates.

3 Experimental results

We predicted the functions of genes of the unicellular eukaryote *S. cerevisiae* at genome and ontology-wide level using the *FunCat* taxonomy [6] and 7 biomolecular data sets, whose characteristics are summarized in Tab. 1.

Table 1. Data sets

Data set	Description	num. of genes	num. of features	num. of classes
Pfam-1	protein domain binary data from *Pfam*	3529	4950	211
Pfam-2	protein domain log E data from *Pfam*	3529	5724	211
Phylo	phylogenetic data	2445	24	187
Expr	gene expression data	4532	250	230
PPI-BG	PPI data from *BioGRID*	4531	5367	232
PPI-VM	PPI data from von Mering experiments	2338	2559	177
SP-sim	Sequence pairwise similarity data	3527	6349	211

Pfam-1 data are represented as binary vectors: each feature registers the presence or absence of 4,950 protein domains obtained from the *Pfam* (Protein families) database [9]. Moreover, we also used an enriched representation of Pfam domains (Pfam-2) by replacing the binary scoring with log E-values obtained with the HMMER software toolkit [10]. The features of the phylogenetic data (Phylo) are the negative logarithm of the lowest E-value reported by BLAST version 2.0 in a search against a complete genome in 24 organisms [11]. The "Expr" data set merges the experiments of Spellman et al. (gene expression measures relative to 77 conditions) [12] with the transcriptional responses of yeast to environmental stress (173 conditions) by Gasch et al. [13]. Protein-protein interaction data (PPI-BG) have been downloaded from the *BioGRID* database, that collects PPI data from both high-throughput studies and conventional focused studies [14]. Data are binary: they represent the presence or absence of protein-protein interactions. We used also another data set of protein-protein interactions (PPI-VM) that collects binary protein-protein interaction data from yeast two-hybrid assay, mass-spectrometry of purified complexes, correlated mRNA expression and genetic interactions [15]. These data are binary too. The "SP-sim" data set contains pairwise similarities between yeast genes represented by Smith and Waterman log-E values between all pairs of yeast sequences [16].

In order to get a not too small set of positive examples for training, for each data set we selected only the FunCat-annotated genes and the classes with at least 20 positive examples. As negative examples we selected for each node/class all genes not annotated to that node/class, but annotated to its parent class. From the data sets we also removed uninformative features (e.g., features with the same value for all the available examples).

We used gaussian SVMs with probabilistic output [17] as base learners. Given a set $\widehat{p}_1, \ldots, \widehat{p}_N$ of trained estimates, we compared on these estimates the results of HTD-CS and HBAYES-CS ensembles with HTD (the cost-insensitive version of HTD-CS, obtained by setting $\tau = 1/2$) and FLAT (each classifier outputs its prediction disregarding the taxonomy). For HTD-CS we set the decision threshold τ by internal cross-validation of the F-measure with training data, while for HBAYES-CS we set the cost factor α to 5 in all experiments. This value provides a reasonable trade-off between between positive and negative examples, as shown by the plots in Figure 2. We compared the different ensemble methods using

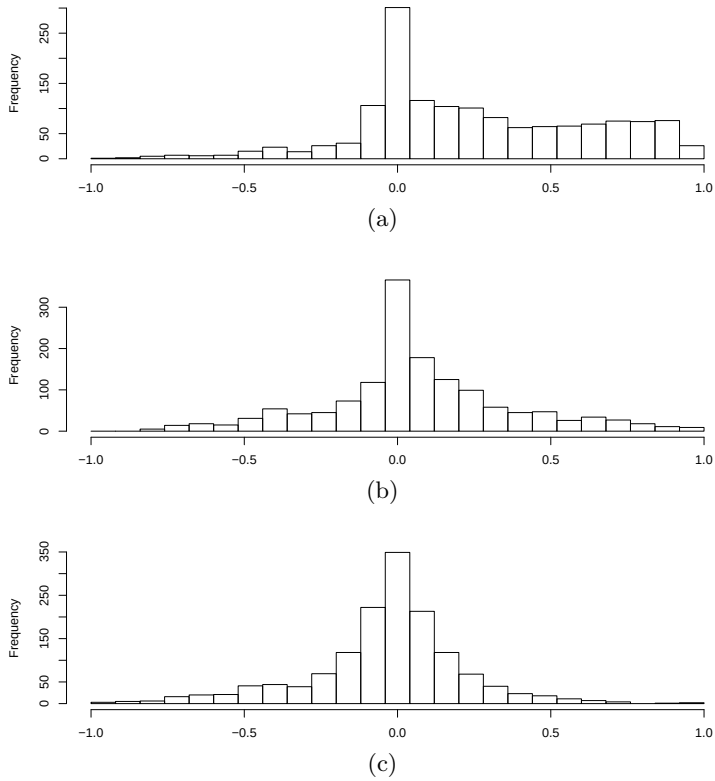


Fig. 1. Histograms of the distribution of the normalized differences between F-measures across FunCat classes and data sets. (a) HBAYES-CS vs. FLAT ensembles; (b) HBAYES-CS vs. HTD ensembles; (c) HBAYES-CS vs. HTD-CS ensembles.

external 5-fold cross-validation (thus without using test set data to tune the hyper-parameters).

For the first set of experiments we used the classical F-score to aggregate precision and recall for each class of the hierarchy. Figure 1 shows the distribution, across all the classes of the taxonomy and the data sets, of the normalized differences $\frac{F_{\text{Bayes}} - F_{\text{ens}}}{\max(F_{\text{Bayes}}, F_{\text{ens}})}$ between the F-measure of HBAYES-CS and the F-measure of each one of the other ensemble methods. The shape of the distribution offers a synthetic visual clue of the comparative performances of the ensembles: values larger than 0 denote better results for HBAYES-CS. In Figure 1.(a) we can observe that HBAYES-CS largely outperforms FLAT, since most of the values are cumulated on the right part of the distribution. The comparison with HTD, Figure 1.(b), shows that HBAYES-CS on average improves on HTD, while essentially a tie is observed with HTD-CS —Figure 1.(c). Indeed the average F-measure across classes and data sets is 0.13 with FLAT ensembles, 0.18 with HTD and 0.22 and 0.23, respectively, with HBAYES-CS and HTD-CS ensembles.

Table 2. Left: Hierarchical F-measure comparison between HTD, HTD-CS, and HBAYES-CS ensembles. Right: win-tie-loss between the different hierarchical methods according to the 5-fold cross-validated paired t-test at 0.01 significance level.

Methods	Data sets							
	Pfam-1	Pfam-2	Phylo	Expr	PPI-BG	PPI-VM	SP-sim	Average
HTD	0.3771	0.0089	0.2547	0.2270	0.1521	0.4169	0.3370	0.2533
HTD-CS	0.4248	0.2039	0.3008	0.2572	0.3075	0.4593	0.4224	0.3394
HBAYES-CS	0.4518	0.2030	0.2682	0.2555	0.2920	0.4329	0.4542	0.3368

win-tie-loss		
Methods	HTD-CS	HTD
HBAYES-CS	2-4-1	6-1-0
HTD-CS	-	7-0-0

In order to better capture the hierarchical and sparse nature of the gene function prediction problem we also applied the *hierarchical F-measure*, expressing in a synthetic way the effectiveness of the structured hierarchical prediction [18]. In brief, viewing a multilabel as a set of paths, hierarchical precision measures the average fraction of each predicted path that is covered by some true path for that gene. Conversely, hierarchical recall measures the average fraction of each true path that is covered by some predicted path for that gene. Table 2 shows that the proposed hierarchical cost-sensitive ensembles outperform the cost-insensitive HTD approach. In particular, win-tie-loss summary results (according to the 5-fold cross-validated paired t-test [19] at 0.01 significance level) show that the hierarchical F-scores achieved by HBAYES-CS and HTD-CS are significantly higher than those obtained by HTD ensembles, while ties prevail in the comparison between HBAYES-CS and HTD-CS (more precisely 2 wins, 4 ties and 1 loss in favour of HBAYES-CS, Table 2, right-hand side). FLAT ensembles results with the hierarchical F-measure are not shown because they are significantly worse than those obtained with any other hierarchical method evaluated in these experiments.

Table 3 shows the per level F-measure results with Pfam-1 protein domain data and Pairwise sequence similarity data (SP-sim). Level 1 refers to the root

Table 3. Per level precision, recall, F-measure and accuracy comparison between FLAT, top-down (HTD), hierarchical top-down cost sensitive (HTD-CS), and hierarchical Bayesian cost sensitive (HBAYES-CS) ensembles. Top: Pfam protein domain data. Bottom: Pairwise sequence similarity data.

Pfam Protein domain																				
FLAT					HTD					HTD-CS					HBAYES-CS					
L.	Prec.	Rec.	F	Acc.	L.	Prec.	Rec.	F	Acc.	L.	Prec.	Rec.	F	Acc.	L.	Prec.	Rec.	F	Acc.	
1	0.76	0.31	0.43	0.88	1	0.76	0.31	0.43	0.88	1	0.66	0.37	0.47	0.88	1	0.74	0.35	0.47	0.89	
2	0.40	0.47	0.35	0.80	2	0.69	0.29	0.39	0.95	2	0.61	0.35	0.43	0.95	2	0.65	0.33	0.43	0.96	
3	0.31	0.46	0.27	0.77	3	0.62	0.25	0.35	0.97	3	0.55	0.30	0.38	0.97	3	0.58	0.30	0.38	0.98	
4	0.15	0.63	0.15	0.54	4	0.56	0.23	0.31	0.98	4	0.53	0.27	0.35	0.98	4	0.54	0.27	0.34	0.98	
5	0.15	0.38	0.17	0.85	5	0.47	0.20	0.27	0.99	5	0.46	0.22	0.29	0.99	5	0.45	0.20	0.26	0.99	
Sequence similarity																				
FLAT					HTD					HTD-CS					HBAYES-CS					
L.	Prec.	Rec.	F	Acc.	L.	Prec.	Rec.	F	Acc.	L.	Prec.	Rec.	F	Acc.	L.	Prec.	Rec.	F	Acc.	
1	0.55	0.41	0.47	0.87	1	0.55	0.41	0.47	0.87	1	0.42	0.58	0.49	0.83	1	0.44	0.56	0.49	0.85	
2	0.08	0.34	0.11	0.74	2	0.30	0.17	0.21	0.94	2	0.24	0.42	0.30	0.90	2	0.27	0.42	0.32	0.92	
3	0.03	0.29	0.05	0.73	3	0.23	0.09	0.12	0.97	3	0.13	0.32	0.18	0.93	3	0.19	0.25	0.20	0.96	
4	0.02	0.49	0.03	0.52	4	0.21	0.07	0.09	0.97	4	0.10	0.37	0.15	0.92	4	0.15	0.18	0.14	0.96	
5	0.01	0.29	0.01	0.68	5	0.04	0.03	0.03	0.98	5	0.05	0.29	0.08	0.94	5	0.10	0.07	0.05	0.98	

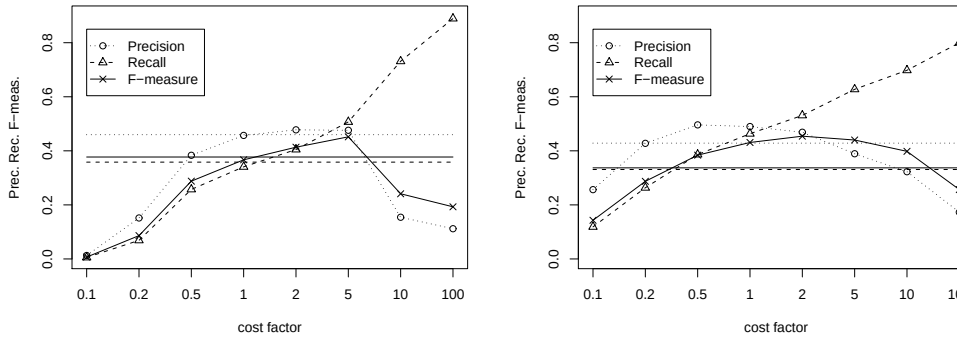


Fig. 2. Hierarchical precision, recall and F-measure as a function of the cost modulator factor in HBAYES-CS ensembles. Left: Protein domain data (Pfam-1). Right: Pairwise sequence similarity data (SP-sim). Horizontal lines refer to hierarchical precision, recall and F-score of HTD ensembles.

nodes of the FunCat hierarchy, level i, $2 \leq i \leq 5$, to nodes at depth i. We can observe that FLAT ensembles tend to have the highest recall, HTD the highest precision, while HBAYES-CS and HTD-CS tend to stay in the middle with respect to both the recall and precision, thus achieving the best F-measure at each level.

The precision/recall characteristics of HBAYES-CS ensemble can be tuned via a single global parameter, the cost factor $\alpha = c_i^- / c_i^+$ (Sect. 2). By setting $\alpha = 1$ we obtain the original version of the hierarchical Bayesian ensemble and by incrementing α we introduce progressively lower costs for positive predictions, thus encouraging the ensemble to make positive predictions. Indeed, by incrementing the cost factor, the recall of the ensemble tends to increase (Fig. 2). The behaviour of the precision is more complex: it tends to increase and then to decrease after achieving a maximum. Quite interestingly, the maximum of the hierarchical F-measure is achieved for values of α between 2 and 5 not only for the two data sets reported in Figure 2, but also for all the considered data sets (data not shown).

The improvement in performance of HBAYES-CS w.r.t. to HTD ensembles has a twofold explanation: the bottom-up approach permits the uncertainty in the decisions of the lower-level classifiers to be propagated across the network, and the cost sensitive setting allows to favor positive or negative decisions according to the value of cost factor. In all cases, a hierarchical approach (cost-sensitive or not) tends to achieve significantly higher precision than a flat approach, while cost-sensitive hierarchical methods are able to obtain a better recall at each level of the hierarchy, without a consistent loss in precision w.r.t. HTD methods — Table 3. We can note for all the hierarchical algorithms a degradation of both precision and recall (and as a consequence of the F-measure) by descending the levels of the trees (Table 3). This fact could be at least in part due to the lack of annotations at the lowest levels of the hierarchy, where we may

have several genes with unannotated specific functions. Despite the fact that the overall performances of HBAYES-CS and HTD-CS are comparable, we can note that HBAYES-CS achieves a better precision (Tab. 3). This is of paramount importance in real applications, when we need to reduce the costs of the biological validation of new gene functions discovered through computational methods. Finally, it is worth noting that the accuracy is high at each level (at least with hierarchical ensemble methods), but these results are not significant, considering the large unbalance between positive and negative genes for each functional class.

4 Conclusions

The experimental results show that the prediction of gene functions needs a hierarchical approach, confirming previous recently published findings [5, 2]. Our proposed hierarchical methods, by exploiting the hierarchical relationships between classes, significantly improve on "flat" methods. Moreover, by introducing a cost-sensitive parameter, we are able to increase the hierarchical F-score with respect to the cost-insensitive version HTD. We observed that the precision/recall characteristics of HBAYES-CS can be tuned by modulating a single global parameter, the cost factor, according to the experimental needs. On the other hand, on our data sets the Bayesian ensemble HBAYES-CS did not exhibit a significant advantage over the simpler cost-sensitive top-down ensemble HTD-CS (see Fig. 1 and Tab. 2). We conjecture this might be due to the excessive noise in the annotations at lower levels of the hierarchy. It remains an open problem to devise ensemble methods whose hierarchical performance is consistently better than top-down approaches even on highly noisy data sets.

In our experiments we used only one type of data for each classification task, but it is easy to use state-of-the-art data integration methods to significantly improve the performance of HBAYES-CS. Indeed, for each node/class of the tree we may substitute the classifier trained on a specific type of biomolecular data with a classifier trained on concatenated vectors of different data [5], or trained on a (weighted) sum of kernels [20], or with an ensemble of learners each trained on a different type of data [21]. This is the subject of our planned future research.

Acknowledgments

We would like to thank the anonymous reviewers for their comments and suggestions. This work was partially supported by the PASCAL2 Network of Excellence (EC grant no. 216886).

References

1. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. Nature Genet. **25** (2000) 25–29
2. Obozinski, G., Lanckriet, G., Grant, C., M., J., Noble, W.: Consistent probabilistic output for protein function prediction. Genome Biology **9** (2008)

3. Sokolov, A., Ben-Hur, A.: A structured-outputs method for prediction of protein function. In: MLSB08, the Second International Workshop on Machine Learning in Systems Biology. (2008)
4. Astikainen, K., Holm, L., Pitkanen, E., Szedmak, S., Rousu, J.: Towards structured output prediction of enzyme function. BMC Proceedings **2** (2008)
5. Guan, Y., Myers, C., Hess, D., Barutcuoglu, Z., Caudy, A., Troyanskaya, O.: Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology **9** (2008)
6. Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., Mewes, H.: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research **32** (2004) 5539–5545
7. Cesa-Bianchi, N., Gentile, C., Tironi, A., Zaniboni, L.: Incremental algorithms for hierarchical classification. In: Advances in Neural Information Processing Systems. Volume 17., MIT Press (2005) 233–240
8. Cesa-Bianchi, N., Gentile, C., Zaniboni, L.: Hierarchical classification: Combining Bayes with SVM. In: Proc. of the 23rd Int. Conf. on Machine Learning, ACM Press (2006) 177–184
9. Finn, R., Tate, J., Mistry, J., Coggill, P., Sammut, J., Hotz, H., Ceric, G., Forslund, K., Eddy, S., Sonnhammer, E., Bateman, A.: The Pfam protein families database. Nucleic Acids Research **36** (2008) D281–D288
10. Eddy, S.: Profile hidden markov models. Bioinformatics **14** (1998) 755–763
11. Pavlidis, P., Weston, J., Cai, J., Noble, W.: Learning gene functional classification from multiple data. J. Comput. Biol. **9** (2002) 401–411
12. Spellman, P., et al.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomices cerevisiae by microarray hybridization. Mol. Biol. Cell **9** (1998) 3273–3297
13. Gasch, P., et al.: Genomic expression programs in the response of yeast cells to environmental changes. Mol.Biol.Cell **11** (2000) 4241–4257
14. Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. **34** (2006) D535–D539
15. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417** (2002) 399–403
16. Lanckriet, G., Gert, R.G., Deng, M., Cristianini, N., Jordan, M., Noble, W.: Kernel-based data fusion and its application to protein function prediction in yeast. In: Proceedings of the Pacific Symposium on Biocomputing. (2004) 300–311
17. Lin, H., Lin, C., Weng, R.: A note on Platt's probabilistic outputs for support vector machines. Machine Learning **68** (2007) 267–276
18. Verspoor, K., Cohn, J., Mnizewski, S., Joslyn, C.: A categorization approach to automated ontological function annotation. Protein Science **15** (2006) 1544–1549
19. Dietterich, T.: Approximate statistical test for comparing supervised classification learning algorithms. Neural Computation **10** (1998) 1895–1924
20. Lanckriet, G., De Bie, T., Cristianini, N., Jordan, M., Noble, W.: A statistical framework for genomic data fusion. Bioinformatics **20** (2004) 2626–2635
21. Re, M, Valentini, G,: Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. JMLR: MLSB 09, 3rd International Workshop on Machine Learning in Systems Biology (2009)

Evaluation of methods in GA studies: yet another case for Bayesian networks

G. Hullám[1], P. Antal[1], Cs. Szalai[2], and A. Falus[3]

[1] Dept. of Measurement and Information Systems, Budapest Univ. of Tech.
[2] Inflammation Biology and Immunogenomics Res. Group, Hung. Acad. of Sci.
[3] Dept. of Genetics, Cell- and Immunobiology, Semmelweis University, Hungary
antal@mit.bme.hu

Abstract. In a typical Genetic Association Study(GAS) several hundreds to thousands of genomic variables are measured and tested for association with a given set of a phenotypic variables (e.g. a given disease state or a complete expression profile), with the aim of identifying the genotypic background of complex, multifactorial diseases. These highly varying requirements resulted in a number of different statistical tools applying different approaches either bayesian or non-bayesian, model-based or model-free. In this paper we evaluate dedicated GAS tools and general purpose feature subset selection(FSS) tools including our own Bayesian model-based tool *BMLA* in a GAS context. In the evaluation we used an artificial data set generated from a reference model with 113 genotypic variables that was based on a real-world genotype data.

1 Introduction

The research on genomic variability received much attention in the past years as one of the most promising areas of genetics research, and several tools were created to aid GAS analysis, particularly the discovery of gene-gene and gene-environment interactions (for an overview see [7]).Earlier multivariate methods designed to detect associations between genotypic variables and the target variable in GAS include *MDR* (Multifactor Dimensionality Reduction [11]), a non-parametric and genetic model-free data mining method, which can also be used in conjunction with several filters such as *ReliefF* [23, 18], *BEAM* (Bayesian Epistasis Association Mapping [27]), which computes the posterior probability that each marker set is associated with the disease via a Markov chain Monte Carlo method, and *BIMBAM*(Bayesian IMputation-Based Association Mapping) which is based on the calculation of Bayes factors [24].

In this paper we compare the performance of these methods and our previously introduced Bayesian network based method in a typical GAS context assuming that the primary goal is (1) the analysis of the relevance of input variables (e.g. SNPs) w.r.t. the target variable (e.g. an indicator of a certain disease); and (2) the exploration of the interdependencies of relevant variables. Note that there are other applicable methods such as *PIA* (Polymorphism Interaction Analysis) [16], and an interaction search method based on external, a priori networks [8] that were not included in this comparative study.

Earlier, we presented the methodology of the Bayesian Multilevel Analysis (BMLA) of the relevance of input variables in [4, 17]. BMLA enables the analysis of relevance at different abstraction levels: model-based pairwise relevance, relevance of variable sets, and interaction models of relevant variables. In the Bayesian model averaging framework each of these levels correspond to a structural property of Bayesian networks (i.e. Markov Blanket Memberships, Markov Blanket sets, and Markov Blanket graphs respectively), and the essence of BMLA is that the estimated posteriors of these properties can be used to assess the relevance of input variables. Furthermore, Markov blanket graph posteriors provide principled confidence measures for multivariate variable selection and facilitates the identification of interaction models of relevant variables (for an extension with scalable structural properties see [5]).

Due to its direct semantics, the Bayesian approach has an in-built automated correction for the multiple testing problem (i.e. the posterior is less peaked with increasing model complexity and decreasing sample size) compared to the hypothesis testing framework.From another point of view, the Bayesian statistical framework is ideal for trading sample complexity for computational complexity (i.e. applying computation intensive model-averaging to quantify the sufficiency of the data). Bayesian conditional methods e.g. using logistic regression or multilayer perceptrons, are widely used in biomedicine and in GASs (e.g., see [3, 13, 6, 22, 19]). Although the conditional approach is capable for multivariate analysis and also copes with conditional relevance and interactions, the model-based approach offers many advantages such as listed below.

1. *Strong relevance.* Clear semantics for the explicit, faithful representation of strongly relevant (e.g. non-transitive) relations (cf. associations)
2. *Structure posterior.* In case of complete data the parameters can be analytically marginalized.
3. *Independence map and causal structure.* It offers a graphical representation for the dependence-independence structure, (e.g. about interactions and conditional relevance) and optionally for the causal relations [21, 9].
4. *Multiple-targets.* It is applicable for multiple target variables [5].

We investigated several probabilistic domain models with promising results [4, 5], relying on these properties of our model-based framework.

2 Probabilistic concepts for GAS

Despite the centrality of "associations" in GASs the refinements of this concept are hardly gaining acceptance in biomedicine, such as strong and weak relevance (cf. non-transitivity and redundancy), conditional relevance (cf. pure interaction), contextual relevance, multivariate relevance (cf. epistasis, complete interaction model, haplotype-level association) or causal relevance. In the following paragraphs, we provide a partial overview on these probabilistic concepts, for a detailed description see [5].

We start with a pure probabilistic definition of relevance, which is defined in a model-free, method-free, cost-free and data-free way [12].

Definition 1 (Relevance) *A feature (stochastic variable) X_i is strongly relevant to Y, if there exists some x_i, y and $s_i = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for which $p(x_i, s_i) > 0$ such that $p(y|x_i, s_i) \neq p(y|s_i)$. A feature X_i is weakly relevant, if it is not strongly relevant, and there exists a subset of features S'_i of S_i for which there exists some x_i, y and s'_i for which $p(x_i, s'_i) > 0$ such that $p(y|x_i, s'_i) \neq p(y|s'_i)$. A feature is relevant, if it is either weakly or strongly relevant; otherwise it is irrelevant.*

A probabilistic definition of relevance can also be given for a set of variables $\mathbf{X}'$ based on the concept of Markov blanket [20].

Definition 1 (Markov boundary). *A set of variables $\mathbf{X}' \subseteq \boldsymbol{V}$ is called a* Markov blanket *set of X_i w.r.t. the distribution $p(\boldsymbol{V})$, if $(X_i \perp\!\!\!\perp V \setminus \mathbf{X}'|\mathbf{X}')_p$, where $\perp\!\!\!\perp$ denotes conditional independence. A minimal Markov blanket is called* Markov boundary. *Its indicator function is denoted by* $\mathrm{MBS}_p(X_i, \mathbf{X}')$.

For the representation of probabilistic relevance, Bayesian networks (BNs) are an adequate choice, since their structural properties are capable of serving such a purpose [20]. They even allow the unambiguous BN representation of relevant variables under a sufficient condition defined in Theorem 1.

Theorem 1. *For a distribution p defined by Bayesian network (G, θ) the variables* $\mathrm{bd}(Y, G)$ *form a Markov blanket of Y, where* $\mathrm{bd}(Y, G)$ *denotes the set of parents, children and the children's other parents for Y [20]. If the distribution p is stable w.r.t. the DAG G, then* $\mathrm{bd}(Y, G)$ *forms a unique and minimal Markov blanket of Y,* $\mathrm{MBS}_p(Y)$ *and $X_i \in \mathrm{MBS}_p(Y)$ iff X_i is strongly relevant [26].*

The induced (symmetric) pairwise relation $\mathrm{MBM}(Y, X_j, G)$ w.r.t. G between Y and X_j is called *Markov blanket membership*. $\mathrm{MBM}(Y, X_j, G)$ indicates whether X_j is in $\mathrm{bd}(Y, G)$ (i.e. X_j is an element of the Markov blanket set of Y).

To include interaction terms into the dependency model of a given variable we proposed the use of the Markov Blanket Graph (MBG) property, a.k.a. classification subgraph [1, 4].

Definition 2 (Markov Blanket Graph). *A subgraph of Bayesian network structure G is called the Markov Blanket Graph or Mechanism Boundary Graph* $\mathrm{MBG}(Y, G)$ *of variable Y if it includes the nodes in the Markov blanket defined by* $\mathrm{bd}(Y, G)$ *and the incoming edges into Y and into its children.*

Finally, note that the definition of conditional relevance corresponds to the concept of pure interaction.

Definition 3 (Conditional Relevance). *Assume that $\boldsymbol{X} = \boldsymbol{X'} \cup \boldsymbol{C'}$ is relevant for $\boldsymbol{Y}$, that is $(\boldsymbol{Y} \not\perp\!\!\!\perp (\boldsymbol{X'} \cup \boldsymbol{C'}))$, and $(\boldsymbol{X'} \cap \boldsymbol{C'} = \emptyset)$. We say that $\boldsymbol{X'}$ is conditionally relevant if $(\boldsymbol{X'} \perp\!\!\!\perp \boldsymbol{Y})$, but $(\boldsymbol{X'} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{C'})$.*

3 GAS Tools

To demonstrate the performance of BMLA compared to other available GAS tools we present the results of a comparative study in Section 4. For this purpose we selected two groups of tools that are capable of analyzing case-control type GASs based on SNP measurements. The first group consists of dedicated Gas tools, designed specifically for GAS analysis, and the second group consists of general purpose feature subset selection methods that are applicable in this GAS context. In the following sections we give a short description for the tools dedicated for GAS analysis.

BEAM: Bayesian Epistasis Association Mapping 1.0 [27]: BEAM uses a Bayesian partitioning model to select SNPs associated with a disease (i.e. the target variable) and their interactions, and computes the posterior probability that each set is associated with the disease via a Markov chain Monte Carlo method. *http://www.fas.harvard.edu/ junliu/BEAM*

BIMBAM: Bayesian IMputation-Based Association Mapping 0.99 [24]: BIMBAM computes Bayes Factors for each SNP, and multi Bayes factors for combinations of SNPs under a linear or logistic regression of target variable(s) on SNPs. *http://stephenslab.uchicago.edu/software.html*

Powermarker 3.25 [15]: PowerMarker contains a set of statistical methods for SNP data analysis. It implements traditional statistical methods for population genetic analysis and also some newly developed methods.
http://statgen.ncsu.edu/powermarker

SNPassoc 1.5.8 [10]: SNPassoc is an R package that provides tools for the analysis of whole genome association studies. It allows the identification of SNP-disease associations based on generalized linear models (depending on the selected genetic inheritance model) and the analysis of epistasis.
http://www.creal.cat/jrgonzalez/software.htm

SNPMStat 3.1 [14]: SNPMStat is an association analysis tool for case-control studies. The program performs a standard association analysis and provides estimated odds ratios, standard error estimates, and Armitage trend tests.
http://www.bios.unc.edu/ lin/software/SNPMStat

Furthermore, we also investigated some general purpose feature subset selection tools, which are as follows:

Causal Explorer 1.4 [2]: Causal Explorer is a library of causal discovery algorithms (such as HITON and IAMB) implemented in MatLab. The algorithms are based on Bayesian Network learning theory, and can also be used for variable selection for classification.
http://discover1.mc.vanderbilt.edu/discover/public/causal_explorer

MDR: Multifactor Dimensionality Reduction 2.0.7 [11]: MDR is a nonparametric and genetic model-free data mining method for detecting nonlinear interactions among discrete genetic and environmental variables. The MDR software also implements a couple of feature selection algorithms to aid the selection of relevant variables. *http://www.multifactordimensionalityreduction.org*

4 Results

We demonstrate the capabilities of BMLA and compare its performance with other GAS tools (presented previously) on an artificial data set, which consists of 5000 complete random samples generated from a reference model containing 113 SNPs (genomic variables) and a clinical variable *Asthma*. The reference model was learned from a real data set containing 1117 samples, and the 113 SNPs were selected from the asthma susceptibility region of chromosome 11q13 [25].

The clinical variable *Asthma* served as the target variable, and the aim of the comparative study was to identify all the relevant variables w.r.t. this target variable. There are 11 SNPs in total that are relevant and therefore are part of the MBG of *Asthma* (see Fig. 1).

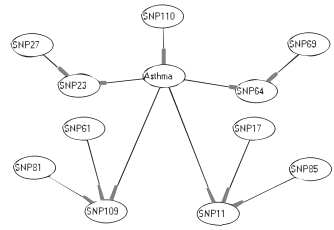


Fig. 1. Markov blanket of the reference model containing all relevant SNPs

Out of the 11 relevant SNPs, 5 are in direct relationship with *Asthma* (i.e. 4 interacting children: SNP11, SNP23, SNP64, SNP109 and 1 single parent: SNP110), and the remaining 6 are interaction terms (SNP17, SNP27, SNP61, SNP69, SNP81, SNP85). The performance of the tools was assessed by comparing their result set of relevant variables against the 11 relevant SNPs of the reference model. In order to measure the effect of varying sample size (i.e. the sufficiency of the data) the computations were run on data sets with sample sizes 500, 1000 and 5000, where the smaller data sets are subsets of larger ones.

Fig. 2 presents the sensitivity for selecting relevant variables for each of the tested dedicated GAS methods. Apart from the overall sensitivity, the sensitivity for identifying relevant variable subgroups (i.e. direct relationships and interactions) is also shown. The result confirms preliminary expectations, that is direct relationships are discovered by almost all of the methods, while interaction terms are ignored by most. Fig. 3 presents the sensitivity measures for the tested general purpose feature subset selection (FSS) methods. The results indicate that the examined FSS methods identify interactions at a significantly higher rate than dedicated GAS tools. Note that due to space limitations only the results gained from the largest data set are shown.

Table 1 shows the sensitivity, specificity and accuracy of the five best performing methods using the complete data set of 5000 samples. Whereas there is

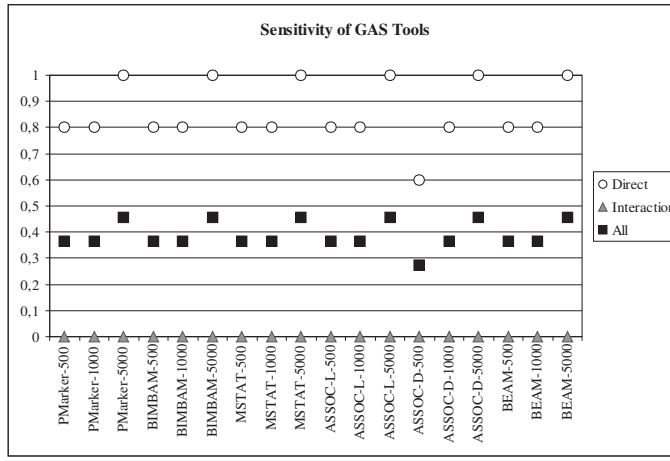


Fig. 2. The performance of dedicated GAS tools: Sensitivity for selecting relevant variables. The figure indicates the sensitivity for identifying all associations, and the two main subtypes separately, i.e. direct relationships and interactions using data sets consisting of 500, 1000 and 5000 samples. Methods are denoted as follows: *PMarker* - PowerMarker, *MSTAT* - SNPMStat, *ASSOC-L* - SNPAssoc using a log-additive inheritance model, *Assoc-D* - SNPAssoc using a dominant inheritance model.

only a slight difference in terms of specificity among the methods, the difference in sensitivity is much more significant.

5 Discussion

The results indicate that the general purpose FSS tools significantly outperformed the tested GAS tools in terms of identifying interactions and conditional relevance.

Basically none of the GAS tools have identified any of the 6 interaction terms of the reference model successfully. Although *BEAM* produced a larger than zero posterior for 3 interaction terms (using 5 chains, 10^6 and $5 * 10^6$ steps for burn-in and length respectively), these were not larger than 0.3 (and thus they were ignored). On the other hand, direct associations were identified by most of the methods correctly. The variation of sensitivity seen on Fig. 2 is due to $SNP110$, which could only be identified from the data set of 5000 samples.

As it can be seen from Table 1, the methods producing the best results all belong to the FSS group. Among them, the best performance was achieved by BMLA using the Cooper-Herskovits (CH) parameter prior. Note that only MBM based results are reported in the paper, because they alone successfully identified the relevant interaction terms of the target variable. However, in several real-world domains the analysis based on MBM probabilities is not enough to identify all relevant variables, and the investigation of MBS properties is re-

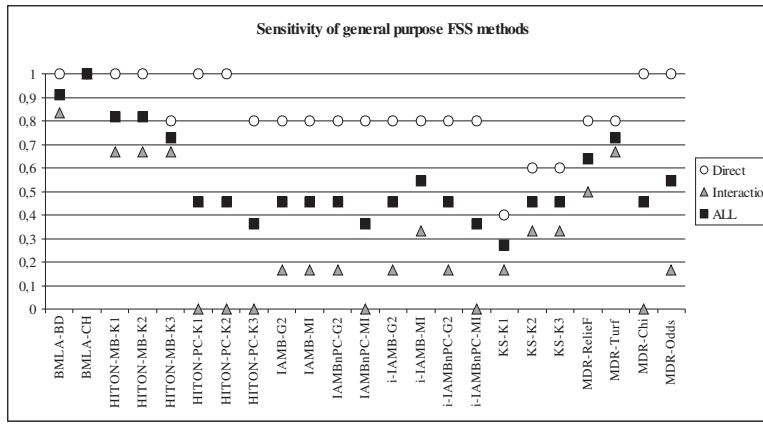


Fig. 3. The performance of general purpose FSS tools: Sensitivity for selecting relevant variables. The suffixes for the methods are as follows: *G2* - based on G^2 statistic, *MI* - based on mutual information, Kn - uses a local test set size of n. Note that Turf, Relief, Chi and Odds denote filters used for variable selection with MDR.

quired. Also note, that though the MBM probabilities are pairwise descriptors, they are generated by Bayesian model averaging, therefore they are normatively multivariate. The second best method was HITON (with several different setups), and the third was MDR in conjunction with its filters ReliefF and TurF. Note that the performance of MDR highly depends on the used filter method, since the exhaustive evaluation of all variables is frequently not feasible. The other FSS methods however identified only a portion of interactions, and missed even some of the direct associations.

On the other hand, the performance of BMLA comes at a high computational cost. On an Intel® Core^{TM}2 CPU 6700 @ 2.66GHz the execution time varied between 5.5 to 7 hours (depending on the data set size) in contrast to all other methods, amongst which the longest execution time was 39.8 minutes. However, with the aid of improving parallelization techniques the execution time of BMLA may significantly be shortened.

6 Conclusion

The presented comparative study has shown, that general purpose FSS tools can be successfully applied in partial genetic association studies and for the purpose of detecting interactions, particularly conditional relevance, among relevant variables, they perform better than dedicated GAS tools. The results also indicated, that BMLA is an adequate choice for evaluating GASs. Its Bayesian network based approach allowed an excellent reconstruction of the reference model, i.e. the identification of relevant variables either in a direct association or in an interaction with the target variable.

Table 1. Sensitivity, specificity and accuracy of the five best performing methods with different parameter settings. The listed methods include BMLA - using two different parameter priors, HITON-MB - with G^2 statistic and a varying local test set size k, MDR - with TurF and RelieF as two pre-filters, interIAMB - based on mutual information(MI), and the Koller-Sahami algorithm(KS).

Method	Sensitivity	Specificity	Accuracy
BMLA-CH MBM	1	0.99	0.9912
BMLA-BD MBM	0.9231	1	0.9912
HITON-MB(G2, k=1)	0.7692	0.98	0.9558
HITON-MB(G2, k=2)	0.7692	0.99	0.9646
HITON-MB(G2, k=3)	0.6923	0.99	0.9558
MDR-TurF	0.6154	0.97	0.9292
MDR-Relief	0.5385	0.96	0.9115
interIAMB(MI)	0.4615	0.96	0.9027
KS(k=3)	0.4615	0.97	0.9115

Finally, note that the capabilities of BMLA can only be fully utilized when the analysis based on MBS or k-MBS structural properties is also carried out [5].

Acknowledgements Supported by grants from the OTKA National Scientific Research Fund (PD-76348), NKTH TECH 08-A1/2-2008-0120 (Genagrid), and the Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences (P.Antal).

References

1. S. Acid, L. M. de Campos, and J. G. Castellano. Learning bayesian network classifiers: searching in a space of partially directed acyclic graphs. *Machine Learning*, 59:213–235, 2005.
2. C.F. Aliferis, I. Tsamardinos, and A. Statnikov. Causal explorer: A probabilistic network learning toolkit for biomedical discovery. In *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '03)*, pages 371–376, 2003.
3. P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial Intelligence in Medicine*, 29:39–60, 2003.
4. P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer. Learning complex bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.
5. P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus. A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. *JMLR Proceeding*, 4:74–89, 2008.
6. D. J. Balding. A tutorial on statistical methods for population association studies. *Nature*, 7:781–91, 2006.
7. H. J. Cordell. Detecting genegene interactions that underlie human diseases. *Nature Reviews: Genetics*, 10(1):392–404, 2009.

8. M. Emily, T. Mailund, J. Hein, L. Schauser, and M. H. Schierup. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, pages 1–10, 2009.
9. C. Glymour and G. F. Cooper. *Computation, Causation, and Discovery.* AAAI Press, 1999.
10. Juan R. Gonzlez, Llus Armengol, Xavier Sol, Elisabet Guin, Josep M. Mercader, Xavier Estivill, and Vctor Moreno. Snpassoc: an r package to perform whole genome association studies. *Bioinformatics*, 23(5):654–655, 2007.
11. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, and White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, 241(2):252–261, 2006.
12. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.
13. C. Kooperberg and I. Ruczinski. Identifying interacting snps using monte carlo logic regression. *Genet Epidemiol*, 28(2):157–170, 2005.
14. D.Y. Lin, Y. Hu, and B.E. Huang. Simple and efficient analysis of disease association with missing genotype data. *American Journal of Human Genetics*, 82(2):444–452, 2008.
15. K. Liu and S. V. Muse. Powermarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, 21(9):2128–2129, 2005.
16. L. E. Mechanic, B. T. Luke, J. E. Goodman, S. J. Chanock, and C. C. Harris. Polymorphism interaction analysis (pia): a method for investigating complex gene-gene interactions. *BMC Bioinformatics*, 9(1):146–158, 2008.
17. A. Millinghoffer, G. Hullám, and P. Antal. On inferring the most probable sentences in bayesian logic. In *Workshop notes on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP-2007)*, pages 13–18, 2007.
18. J. H. Moore and B. C. White. Tuning relieff for genome-wide genetic analysis. In *Lecture Notes in Computer Science: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 166–175. Springer Berlin - Heidelberg, 2007.
19. M. Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2007.
20. J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Francisco, CA, 1988.
21. J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge U. Press, 2000.
22. M. A. Province and I. B. Borecki. Gathering the gold dust: Methods for assessing the aggregate impact of small effect genes in genomic scans. In *Proc. of the Pacific Symposium on Biocomputing (PSB08)*, volume 13, pages 190–200, 2008.
23. M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1):2369, 2003.
24. B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate genes and quantitative traits. *PLoS Genetics*, 3(7):e114, 2007.
25. C. Szalai. Genomic investigation of asthma in human and animal models. In A. Falus, editor, *Immunogenomics and Human Disease*, pages 419–441. Wiley, London, 2005.
26. I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy,filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342, 2003.
27. Y. Zhang and J. S Liu. Bayesian inference of epistatic interactions incase-control studies. *Nature Genetics*, 39(9):1167–1173, 2007.

Evaluation of Signaling Cascades Based on the Weights from Microarray and ChIP-seq Data

Zerrin Isik[1], Volkan Atalay[1], and Rengül Çetin-Atalay[2]

[1] Department of Computer Engineering, Middle East Technical University, Ankara TURKEY
{zerrin.sokmen,volkan}@ceng.metu.edu.tr
[2] Department of Molecular Biology and Genetics, Bilkent University, Ankara TURKEY
rengul@bilkent.edu.tr

Abstract. In this study, we analyzed the combination of the ChIP-seq and the transcriptome data and we integrated these data into signaling cascades. Integration was realized through a framework that was hybrid of data-driven and model-driven approaches. An enrichment model was constructed to evaluate signaling cascades which resulted in specific cellular processes. We used ChIP-seq data and microarray data from public databases which were obtained from HeLa cells under oxidative stress having similar experimental setups. The genes associated with OCT1 transcription factor were identified based on ChIP-seq data. Both ChIP-seq and array data were analyzed by percentile ranking for the sake of simultaneous data integration on specific genes from signaling cascades. Signaling cascades from KEGG pathway database were subsequently scored by taking sum of the individual scores of the genes involved within the cascade and this score information transferred to en route of the signaling cascade to form final score. Furthermore, we evaluated oxidative stress effected cellular processes based on the final scores. We believe that signaling cascade model based framework that we describe in this study is applicable to other transcriptome data analysis.

Key words: evaluation of signaling cascades, chip-seq, gene expression

1 Introduction

Microarray experiments enable researchers to access the transcriptome information related to the state of several thousands of genes under a particular experimental condition. Traditional analysis methods for microarray data, output a list of significant genes specific to the performed experiments. In order to associate the list of genes to a specific cellular process secondary tools and databases are used. Therefore, research focuses on the analysis of the biological pathways rather than individual genes [1]. Several gene prioritization methods attempt to determine the similarity between candidate genes and genes known to play a role in defined biological processes or diseases [2–5]. Therefore, it is clear that available data from multiple sources (e.g. Gene Ontology annotations,

protein domain databases, biological networks, published literature, gene expression data etc.) would enrich the analysis. A variety of methods have been developed to analyze and visualize microarray data by considering known biological networks [6–10]. These methods identify significant functional terms or biological pathways by applying several statistical significance tests. They also overlay gene expression data into known pathways to visualize experiment specific gene regulations [11, 12]. Additionally, these tools apply graph theory and calculate significance scores on the pathways. Nevertheless, these tools depend on the primary significant gene lists.

Chromatin ImmunoPrecipitation (ChIP) combined with genome resequencing (ChIP-seq) technology provides protein DNA interactome data. ChIP-seq technology is expected to be popularly used for the analysis of gene expression signatures, as it happened with microarray technology. Transcription factors (TFs) bind to specific DNA sequences and turn transcription of target genes on or off. In order to explore accurate prediction of pathway activities, ChIP-seq experiments provide detailed knowledge about target genes. ChIP-seq experiments and computational analysis methods in literature have been at initial stages [13]. Although there is a few number of early stage analysis tools for ChIP-seq data, secondary gene annotation methods should also be integrated like in the case of microarray data analysis. Therefore, we considered to integrate ChIP-seq and gene expression experiments to identify target genes responsible of a specific cellular process.

In this study, we analyzed ChIP-seq and gene expression data together by applying computational methods and mapped gene scores to biological signaling cascades. After combining ChIP-seq and gene expression profiles, we constructed an enriched model to evaluate the signaling cascades under the control of specific biological processes (Figure 1).

2 System and Methods

2.1 Data Processing

Experimental data sources of this study were obtained from NCBI GEO database. We selected ChIP-seq and microarray data from GEO datasets (GSE14283, GSE4301). We used ChIP-seq data by Kang et al. which aimed to identify transcription regulation role of OCT1 on HeLa cells under oxidative stress [14]. Raw ChIP-seq data of OCT1 protein contains approximately 3.8 million reads. Initially, significant peak regions in raw data were explored by applying peak detection method of CisGenome tool [15]. Peak detection method scans the genome with a sliding window (w=100, s=25) and identifies regions with read counts greater than 10 reads for significant binding regions. We obtained 5080 putative peak regions over entire human genome. In order to compute significancy of each peak region, we set a percentile rank value for each peak region by considering total number of reads involved in that region.

$$ReadRank(r) = \frac{cf_l + 0.5(f_r)}{T} \tag{1}$$

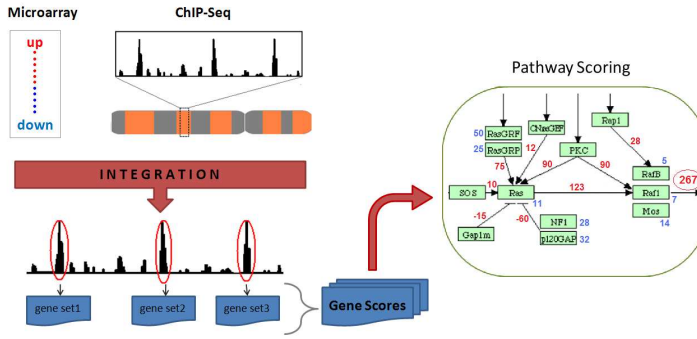


Fig. 1. Process diagram of the proposed method. The integration stage combines ChIP-seq and transcriptome data to obtain scores of genes related with OCT1 transcription factor. In the next stage, signaling cascades activated by transcription of OCT1 are identified by exploring scores of each signaling cascade.

where cf_l is the cumulative frequency for all scores lower than the score of the peak region r, f_r is the frequency of the score of peak region r, and T is the total number of peak regions. $ReadRank(r)$ score ranges from 0 to 1. After identification of OCT1 high quality peak regions, we mapped to the TSS of the genes within a region $\pm$10000 bp. Total number of neighboring genes associated with high quality peak regions was 260.

The microarray data used in the study was also obtained from HeLa cells under oxidative stress condition (Murray expression data) [16]. We calculated fold-change (gene expression log ratio) of two channels (red, green) for control and oxidative stress experiments.

$$FoldChange(x) = log_2(\frac{\overline{ch2_x}}{\overline{ch1_x}}) \tag{2}$$

where $\overline{ch1_x}$ and $\overline{ch2_x}$ represent the mean value of channel 1 and channel 2 of gene x, respectively. We observed that half of the genes have very low fold changes (less than 0.2 fold). In order to assign a rank value of the gene expression, we applied Equation 3 which involves the same computation with $ReadRank$.

$$ExpRank(x) = \frac{cf_l + 0.5(f_x)}{T} \tag{3}$$

where cf_l is the cumulative frequency for all fold-change values lower than the fold-change value of the gene x, f_x is the frequency of the fold-change value of gene x, and T is the total number of genes in microarray chip. If the magnitude of fold change is very close to 0, the rank value is close to 0. Otherwise, rank value of a gene varies between 0 and 1 according to magnitude of its fold change.

2.2 Integration of ChIP-seq and Microarray Data

The gene set extracted from OCT1 ChIP-seq data and Murray expression data for a gene were associated by taking their weighted linear combinations.

$$Score(x) = c_{chip}ReadRank(x) + c_{exp}ExpRank(x) \tag{4}$$

where $ReadRank(x)$ is the ChIP-seq read rank value of gene x given by Equation 1, $ExpRank(x)$ is the expression rank value of gene x indicated by Equation 3, and c_{chip} and c_{exp} are the coefficients of two data sources. In order to consider their effects equally, 0.5 was assigned to both c_{chip} and c_{exp}.

2.3 Scoring of Signaling Cascades

In order to assign scores to signaling cascades which control biological process we used KEGG pathway as the model. For this purpose, we converted selected KEGG pathways into the graph structures by using KGML files. A node of the graph represents gene product, chemical compound or biological process represented by other KEGG pathways. The edges represent the relations (i.e. activation, inhibition) between the nodes. We enumerated each signaling cascade from a specific KGML file that leads to biological process of the selected pathway. If the edge between two nodes is activation, the total score of that node is transferred directly. If the edge is inhibition, the total node score is transferred with a negative value (Figure 2). If a gene, involved in a pathway, has no score, the value of $Score(x)$ was set to zero. In order to consider processing order of the genes in actual pathway map, we performed score computations in the order of cascading nodes. Total score of a signaling cascade was computed by applying that score flow mechanism up to the goal node: biological process. Algorithm 1 describes general steps of the biological score computation.

Total score of a signaling cascade P is computed by taking sum of all possible biological processes under the control of P.

$$Enrichment(P) = \sum_{s=1}^{N} outputScore(s) \tag{5}$$

where $outputScore(s)$ is total path score of the biological process s, N is the total number of biological processes under the control of P. The average score of the signaling cascade P was computed to discover oxidative stress effected signaling cascades and assign a significance score to them.

$$AverageEnrichmentScore(P) = \frac{Enrichment(P)}{N} \tag{6}$$

where $Enrichment(P)$ is the total score of the signaling cascade P, N is the total number of genes involved in that signaling cascade.

Algorithm 1 : Computing Score of Signaling Cascades

Input: Graph $\underline{P}$, has *nodes* and *edges* arrays
Score: indicates self score of each node given by our method
outputScore: contains output edge score of each node
Initialization:
Apply Breadth-First Search algorithm
Extract initialization (ancestor of $\underline{P}$) nodes: $initialNodes$ = start node(s) of $\underline{P}$
$otherNodes = nodes \setminus initialNodes$
Score Computation:
for $i = 1$ to $length(initialNodes)$ **do**
  $outputScore[initialNodes[i]] = Score[initialNodes[i]]$
end for
for $j = 1$ to $length(otherNodes)$ **do**
  $ancestorNodes$ = ancestor node(s) of $otherNodes[j]$
  $outputScore[j] = Score[j]$
  for $k = 1$ to $length(ancestorNodes)$ **do**
    $e = E(k, j)$ {the edge between $ancestorNodes[k]$ and $otherNodes[j]$}
    if type of e is activation **then**
      $sign[k] = 1$ {assign weight of activation edge}
    else
      $sign[k] = -1$ {assign weight of inhibition edge}
    end if
    $outputScore[j] += outputScore[k] * sign[k]$ {sum up weight of incoming edge}
  end for
  if $outputScore[j] < 0$ **then**
    $outputScore[j] = 0$ {negative score is originated by only inhibition edges}
  end if
end for
Output: *outputScore* of outcome biological processes in graph $\underline{P}$.

3 Results and Discussion

3.1 Evaluation of Signaling Cascades

After assigning gene scores by integrating rank scores from ChIP-seq and array data, biological annotation of these gene scores was performed by evaluating several signaling cascades obtained from KEGG database. Mapping gene scores onto pathways can provide the determination of specific regulation motifs driving different responses in several signaling cascades. An example about this pathway enrichment is illustrated in Figure 2. A sub-cascade represented in this figure starts with the initial activation nodes of Jak-STAT signaling cascade and ends with the goal: *Apoptosis* biological process. We attributed gene scores on the nodes and reflected that information to en route of the signaling cascade. Thus, the final biological process *Apoptosis* under the control of Jak-STAT signaling cascade was scored. The total score for Apoptosis biological process computed by oxidative stress expression data (Figure 2-A) was higher than that of the control (Figure 2-B).

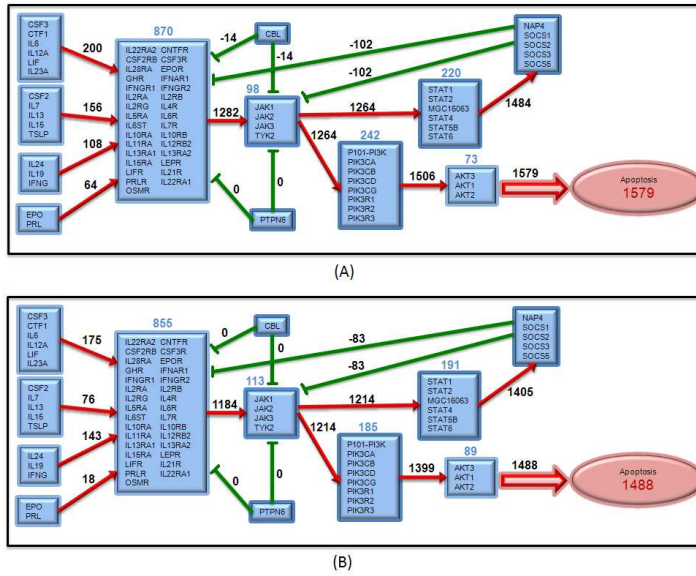


Fig. 2. Assignment of combined ChIP-seq and array scores to a sub-cascade (*Apoptosis*) leading under oxidative stress (A) and control experiment (B). The number on each node (gene) represents self-score of the gene. Edge weights are computed by the sum of node scores. Red and green edges represent activation and inhibition properties, respectively.

In order to highlight the novelties of our model driven framework for transcriptome data analysis we also applied *kegArray* tool [17] to Murray oxidative stress data over Jak-STAT signaling cascade (Figure 3). Several tools, similar to kegArray, map expression data over pathways; however, they could not assign a score to the biological end process. On the other hand, our method provides a quantitative measure to evaluate biological activity of a pathway in a specific process.

We applied our framework to 4 KEGG pathways having 15 signaling cascades: Jak-STAT signaling, TGF-β signaling, Apoptosis, and MAPK signaling pathways. For all pathways, oxidative stress data obtained higher scores with respect to control experiment (Table 1). Control experiment obtained 4, 18, 11, and 20 % lower scores compared to oxidative stress scores in Jak-STAT signaling, TGF-β signaling, Apoptosis, and MAPK signaling pathways respectively. When the average scores of outcome biological processes are compared, *Apoptosis* biological process in Jak-STAT signaling cascade produced the highest score,

224.73, by using oxidative stress experiment. Thus, we can conclude that the most effected biological process under oxidative stress condition and transcription of OCT1 protein is *Apoptosis* biological process.

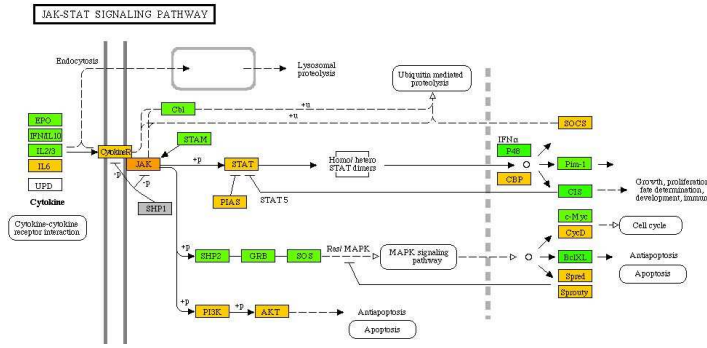


Fig. 3. Mapping of Murray oxidative stress data to Jak-STAT signaling cascade by using kegArray tool. Green and orange color indicate down-regulation and up-regulation values, respectively.

4 Conclusion

In general, current approaches which integrate transcriptome data to molecular pathways are either data driven or model driven. In this study, we applied a hybrid approach which integrates large scale (i.e. transcriptome, ChIP-seq) data to quantitatively assess the weight of a signaling cascade under the control of a biological process. In our framework signaling cascades act as models. Our hybrid approach equally utilizes the signaling cascade intrinsic properties (i.e. edge and node specifications) and scores genes through large scale data.

Pathway enrichment is typically applied on the nodes of the pathways. The enriched pathways can be visualised as nodes and edges so that the user would choose based on observed highlighted nodes. However, in our study we computed the scores of activated processes. We used ChIP-seq data in order to further enrich scores of the specific genes. Therefore, we believe that, if available, ChIP-seq and other large scale data can be further integrated into this framework. We attribute integrated data on the nodes and reflect this information to en route of the pathway as scores. These scores reflect the current activity of analyzed pathway. Our framework in its current state can be applied to directed acyclic graphs. Actually, biological signaling cascades also act in acyclic directed manner, since the signal flow is through membrane to nucleus or vice versa.

KEGG ID	Biological Process	Control Sample		Oxidative Stress	
		Total Score	Average Score	Total Score	Average Score
hsa04630	Apoptosis	5633	217.80	5843	224.73
	Cell Cycle	5447	209.50	5558	213.76
	Ubiquitin mediated proteolysis	2587	99.5	2754	105.92
	MAPK signaling	1336	51.38	1358	52.23
hsa04350	Cell Cycle	158	2.92	166	3.07
	MAPK signaling	44	0.81	76	1.40
	Apoptosis	52	0.96	66	1.22
hsa04210	Survival	2222	37.66	2762	46.81
	Apoptosis	2668	45.22	2709	45.91
	Degradation	1984	33.62	2188	37.08
hsa04010	Proliferation & differentiation	19346	172.73	22315	199.24
	Cell Cycle	2533	22.61	3771	33.66
	Apoptosis	1652	14.75	2949	26.33
	p53 signaling	832	7.42	1135	10.13
	Wnt signaling	185	1.65	288	2.57

Table 1. Assigned scores of hsa04630 Jak-STAT signaling, hsa04350 TGF-β signaling, hsa04210 Apoptosis, hsa04010 MAPK signaling pathways for control and oxidative stress experiments. Total score indicates overall score of each biological process. Total score is divided by the amount of nodes of the analyzed cascade to obtain the average score.

Currently used pathway analysis tools do not assign any roles to genes which are not differentially expressed for the enrichment. However our hybrid approach considers signal relaying molecules even though they are not differentially expressed. We believe that our hybrid approach better represents a biological process rather than ignoring a gene product which is not differentially expressed but present in the biological process.

References

1. Cordero, F., Botta, M., Calogero, R.A.: Microarray data analysis and mining approaches. Brief. in Funct. Genom. and Proteom. 1–17, (2008)
2. Aerts, S., Lambrechts, D., Maity, S., Loo, P.V., Coessens, B., Smet, F.D., Tranchevent, L.C., Moor, B.D., Marynen, P., Hassan, B., Carmeliet, P., Moreau, Y.: Gene prioritization through genomic data fusion. Nat.Biotech. 24, 537–544, (2006)
3. Bie, T.D., Tranchevent, L.C., van Oeffelen, L.M.M., Moreau, Y.: Kernel based data fusion for gene prioritization. Bioinformatics 23 , i125–i132, (2007)
4. Lopez-Bigas, N., Ouzounis, C.A.: Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 32, 3108–3114, (2004)
5. Kent, W., Hsu, F., Karolchik, D., Kuhn, R.M., Clawson, H., Trumbower, H., Haussler, D.: Exploring relationships and mining data with the ucsc gene sorter. Genome Res. 15, 737–741, (2005)

6. Al-Shahrour, F., R. Diaz-Uriarte, and J. Dopazo: Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. Bioinformatics 20(4), 578–580, (2004)
7. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat. Genet., 31, 19–20, (2002)
8. Chung,H.J., Park,C.H., Han,M.R., Lee,S., Ohn,J.H., Kim,J.,Kim,J. and Kim,J.H. ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using scalable Vector Graphics. Nucleic Acids Res., 33, W621–W626, (2005)
9. Mlecnik,B., Scheideler,M., Hackl,H., Hartler,J., Sanchez-Cabo,F. and Trajanoski,Z.PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic Acids Res., 33, W633–W637, (2005)
10. Goffard N. and Weiller G. PathExpress: a web-based tool to identify relevant pathways in gene expression data Nucleic Acids Res. 35, W176–W181, (2007)
11. Ingenuity pathway analysis, http://www.ingenuity.com
12. Nikitin,A., Egorov, S., Daraselia, N. and Mazo, I. : Pathway studio–the analysis and navigation of molecular networks, Bioinformatics, vol. 19, pp. 2155–2157, (2003).
13. Rosenbluth JM, Mays DJ, Pino MF, Tang LJ, Pietenpol JA. A Gene Signature-Based Approach Identifies mTOR as a Regulator of p73, Mol. and Cell. Bio., 28(19), 5951–5964, (2008)
14. Kang J, Gemberling M, Nakamura M, Whitby FG, Handa H, Fairbrother WG, Tantin D. A general mechanism for transcription regulation by Oct1 and Oct4 in response to genotoxic and oxidative stress. Genes Dev. 23(2), 208–222, (2009)
15. Ji H., Jiang H., Ma W., Johnson D.S., Myers R.M., and Wong W.H. An integrated software system for analyzing chip-chip and chip-seq data. Nat Biotechnol., 26(11), 1293–1300, (2008)
16. Murray JI, Whitfield ML, Trinklein ND, Myers RM, Brown PO, Botstein D. Diverse and specific gene expression responses to stresses in cultured human cells. Mol Biol Cel, 15(5), 2361–2374, (2004)
17. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34:D354–7, (2006).

Matching Models to Data in Modelling Morphogen Diffusion

Wei Liu and Mahesan Niranjan

School of Electronics and Computer Science, University of Southampton
Southampton, SO17 1BJ, United Kingdom
{wl08r,mn}@ecs.soton.ac.uk

Abstract. The mechanism by which spatial patterns are established during embryonic development is usually modelled as passive diffusion of morphogen proteins translated from maternally deposited messenger RNAs. Such diffusion models assume a constant supply of morphogens at the source throughout the establishment of the required profile at steady state. Working with the bicoid morphogen which establishes the anterior-posterior axis in the Drosophila embryo, we note that this constant source assumption is unrealistic since the maternal mRNA is known to decay after a certain time since egg laying. We numerically solve the reaction diffusion equation for one dimensional morphogen propagation and match the resulting solution to measured data. By minimising the squared error between model outputs and measurements published in the FlyEx database, we show how parameters of diffusion rate, mRNA and protein decay constants, and the onset of maternal mRNA decay can be assigned sensible values.

Key words: Morphogen diffusion, mRNA degradation, Parameters estimation

1 Introduction

Passive diffusion of a class of molecules known as morphogens as a mechanism that helps to establish spatial patterns of gene expression during embryonic development was proposed several decades ago by Turing [1]. Several morphogen molecules have been discovered since his postulation. In the fruit fly *Drosophila melanagaster*, the maternally deposited gene Bicoid establishes the first morphogen gradient along the anterior posterior axis. Properties of the steady state profile set up by such diffusion include precision of a decision threshold in the presence of variabilities in the form of embryo length, amount of maternally deposited mRNA etc. Several authors have analysed such formation of developmental precision [2, 3].

Bergmann *et al.* [4] have analysed properties of the spatio-temporal morphogen profile prior to the establishment of a steady state. They argue that desirable properties of the steady state profile are also realised in the pre-steady

state region, thereby enabling reliable downstream gene expression early in the nuclear cleavage cycles.

While such passive diffusion is a widely of how a concentration gradient is established, the topic may have to be re-visited in the light of recent experimental findings. Hecht *et al.* [5] offer an alternate model based on cytoplasmic flow, motivated by the argument that the quantitative properties of the morphogen profiles established require higher values of diffusion constant than have been experimentally measured recently [6].

In this paper we consider an enhancement to the passive one dimensional diffusion model, hitherto ignored in the literature. We incorporate a more realistic model of the morphogen source. In classical analysis using the diffusion model, the source is assumed to provide a constant supply of morphogen from the beginning until a steady state profile is established at cleavage cycle 14, which takes place about 130 mins from the laying of the egg. This is an unrealistic assumption, since the maternal mRNA should be expected to decay. While literature evidence on this is not quantitative, there is suggestion, for example see [7], that the maternal bicoid starts to decay rapidly from cleavage cycle 12 onwards. In this work we explicitly model the source as a constant supply followed by exponential decay and compute the solution to the reaction diffusion system. In Flyex Database [8], bicoid integrated data in nuclear cleavage cycle $14A$ in one- dimension is used as measured data. Cycle $14A$ is approximately 50 mins in duration and is divided into 8 equal temporal classes of 6.5 mins duration [9]. By matching the resulting profile to available data from the FlyEx database [8], during cleavage cycle 14, we estimate parameters of the diffusion model, including the point in time maternal mRNA decay begins. By incorporating a realistic source model, our analysis completes the passive model and sets the framework for parameter estimation including uncertainties, as more data becomes available. An important finding of the study is that matching model output to data in the post-peak profile. We recover parameter values typically recommended by other authors for desirable profiles in the steady and pre-steady state profiles, which, to some extent, is a validation of our model.

2 Model and Implementation

The reaction diffusion equation used to model morphogen is given by

$$\frac{\partial}{\partial t}M(x,t) \;=\; D\,\frac{\partial^2}{\partial x^2}M(x,t) \;-\; \tau_p^{-1}\,M(x,t) \;+\; S(x,t),$$

where $M(x,t)$ is the morphogen concentration as a function of space and time. D, the diffusion constant, τ_p, the half-life of the morphogen protein and $S(x,t)$, the source at the anterior end. The usual assumption in solving this model is that the sourse is constant:

$$S_{con} = S_0\delta(x)\Theta(t)$$

where S_0 is the production rate, $\delta(x)$ is the Kronecker delta function and $\Theta(t)$ is Heaviside step function. Thus we have a point source at $x = 0$, the anterior pole which is zero for $t < 0$ and has a constant magnitude from time zero onwards. This equation can be easily solved analytically using the method proposed in [4].

Here, we work with a source model which has a constant part during which the maternal mRNA is kept stable, followed by an exponentially decaying part, due to maternal mRNA decaying from about cleavage cycle 12 of the developing embryo. This source is modelled as follow:

$$S_{com} \;=\; S_0\delta(x)\;(\Theta(t) - \Theta(t - t_0)) \;+\; S_0\delta(x)\,\Theta(t - t_0)\,\exp\left\{-\frac{t - t_0}{\tau_m}\right\}$$

While the solution to the reaction diffusion equation with such a source may also be tractable analytically, we instead chose the easy option of numerically integrating them with the MATLAB toolbox `pdepe`. We obtain solutions with the constant source up to time t_0, then with an exponentially decaying source starting at t_0, and combined the solutions by linear superposition, taking care to match the boundary conditions.

3 Results and Discussion

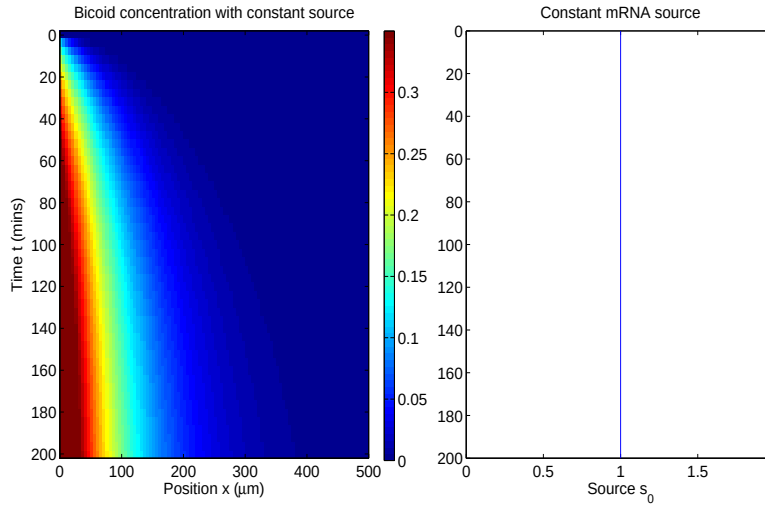


Fig. 1. Widely used models of bicoid diffusion assume a constant source, leading to an exponentially decaying steady state model. The left panel is an intensity profile of morphogen concentrations jointly in time and along the length of the embryo. The right panel shows the constant supply of bicoid protein at the anterior end of the embryo.

Fig. 1 shows the solution to the reaction diffusion model of morphogen propagation, where the source bicoid supply is a constant. This is the widely used model which at steady state sets up an exponential profile. We obtained this solution by numerically integrating the differential equation using `pdepe` Toolbox in MATLAB with $D = 1.7\,\mu m^2/s$ and $\tau_p = 104$ mins. Fig. 2 shows the solution to the diffusion model, which we propose as being more realistic, in which the source is a combination of a constant supply followed by an exponential decay. As expected, the solution to this system, evaluated numerically using the `pdepe` Toolbox, sets up a spatially decaying profile which subsequently decays to zero. For this simulation, D and τ_p are set to the same values as above while the decay rate of maternal bicoid mRNA was set as $\tau_m = 1/3\,\tau_p$. Following [7], mRNA degradation was set to start at 120 mins.

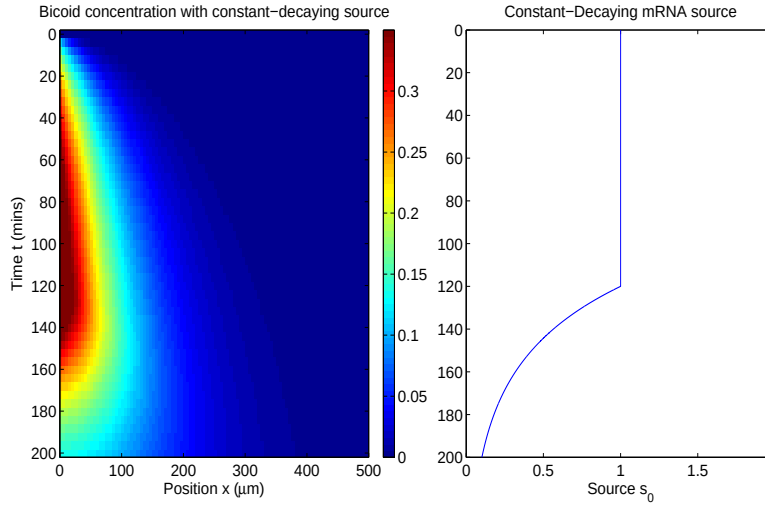
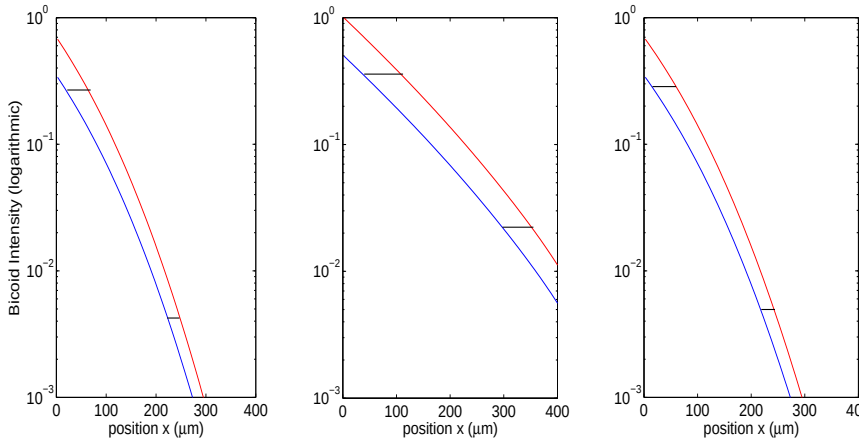


Fig. 2. Space-time concentration of bicoid with the more realistic model (right) that is constant for a certain period of time, followed by exponential decay as the maternally deposited bicoid mRNA is degraded. Precisely when the decay of maternal mRNA begins is a parameter to be optimised.

Bergmann *et al.* [4] have analysed the sensitivity of a morphogen boundary to the source proteins dosage. Fig.3 shows similar boundary shift sensitivities at pre-steady state, peak profile and post-peak profile. We note that properties of the boundary observed in [4], also appear to hold for much larger in time (i.e. post-peak profile).



(a) Pre-steady-state (b) Steady-state (c) Post-peak-state

Fig. 3. Intensity comparison with different source production rate. Original intensities correspond to blue lines while red lines show the intensities with 2-fold increase in production rate. (a) shows bicoid concentration in pre-steady state. (b) is in steady state and (c) is in post-peak state.

3.1 Matching Parameter Values to Data

We used intensity profiles published in the FlyEx database [8] to estimate what sensible parameters of the model were. Throughout this paper we used the squared error between model output and measured intensities to evaluate error in estimating the parameters.

$$E = \sum_{t=T_1}^{T_2} \sum_{x=1}^{L} \{M(x, t) - M_d(x, t)\}^2,$$

where T_1 and T_2 were the boundaries of cleavage cycle $14A$ for which data, $M_d(x,t)$, was available at eight uniformly sampled time points. Cleavage cycle $14A$ was of specific interest, because, it is during this period, that cellularization sets in and the bicoid profile established begins to decay due to the decay of the source mRNA and the diffused protein.

Fig. 4 shows intensities of morphogen output by the model with a constant source followed by an exponentially decaying source, and the measured data from FlyEx. It shows that a good match to the measured data is obtainable in the post-peak stages of morphogen profile, jointly in space and time. To the best of our knowledge, these stages have not attracted interest in the literature, and the popular model with a constant morphogen source is clearly incorrect in these stages.

Fig. 5 shows the variation in modelling error as functions of the four parameters (D,t_0,τ_p and τ_m), where we have held three of the four constant at their best

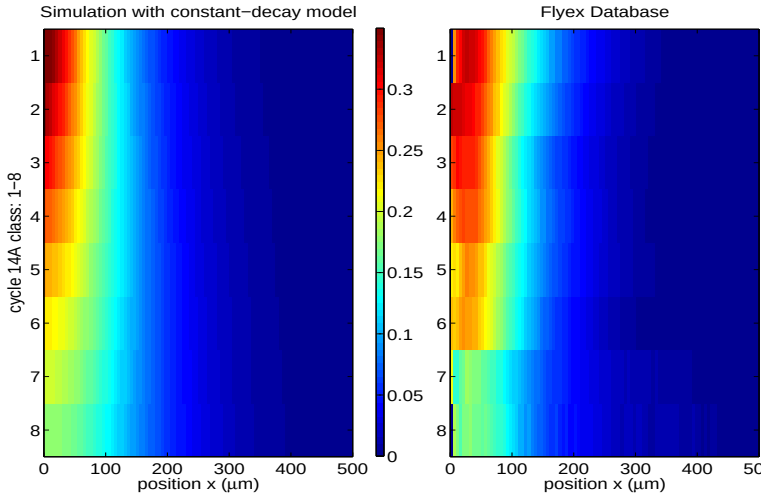


Fig. 4. Comparison of model based and measured data of bicoid intensities

estimates found in the literature and varied the fourth. The resulting error is unimodal with respect to each of the parameters, and the corresponding minima are marked in red on the figures.

3.2 Diffusion Constant

Broadly, the optimum values are in agreement with ranges specified in the literature. Bergmann *et al.* [4] suggest a value for diffusion constant in the range $0.3 \sim 3\ \mu m^2/s$. When matching the post-peak decay phase, as we have done here, the optimum value obtained is $D = 1.8 \mu m^2/s$.

3.3 When does Maternal mRNA begin to decay?

There is very little indication in the literature as to when maternal mRNA begins to decay. Surdej *et al.* [7] suggest that maternal mRNA is stable during first 2 hours and then decays rapidly between 2 and 3 hours after fertilization. Here we find that when mRNA starts to degrade at 120 mins the difference between simulated and measured data is minimum. When degradation starts later than 140 mins or earlier than 120 mins, the error increases significantly. If mRNA decay starts much later, the model tends to be the one with constant source and the error is high.

3.4 Messenger RNA and Protein Half-lives

To the best of our knowledge there aren't reliable published measurements of protein decay rates for bicoid. In order to confer desirable pre-steady state prop-

erties to morphogen profiles, Bergmann *et al.* [4] suggested values for τ_p should be higher than the range $65 \sim 100$ mins. The motivation in their model was the observation that some gap gene expressions were observable prior to the establishment of steady morphogen profiles [10]. Our best estimate of protein decay time, matching in the post-peak region is 111 mins, and is in agreement with what was required for desirable pre-steady state properties.

Our estimate of bicoid mRNA half-life is 29 mins, which is nearly 1/3 of bicoid proteins decaying time from our model. This satisfies the observation that, in general mRNA decays much faster than the corresponding protein. [7] suggest a decay time constant of less than 30 mins, but with no experimental evidence to support it.

Fig.6 shows the variation in modelling error as a function jointly in two of the variables: diffusion constant and onset of maternal mRNA decay. This, too, is unimodal and achieves a minimum in the range of sensible values as discussed earlier.

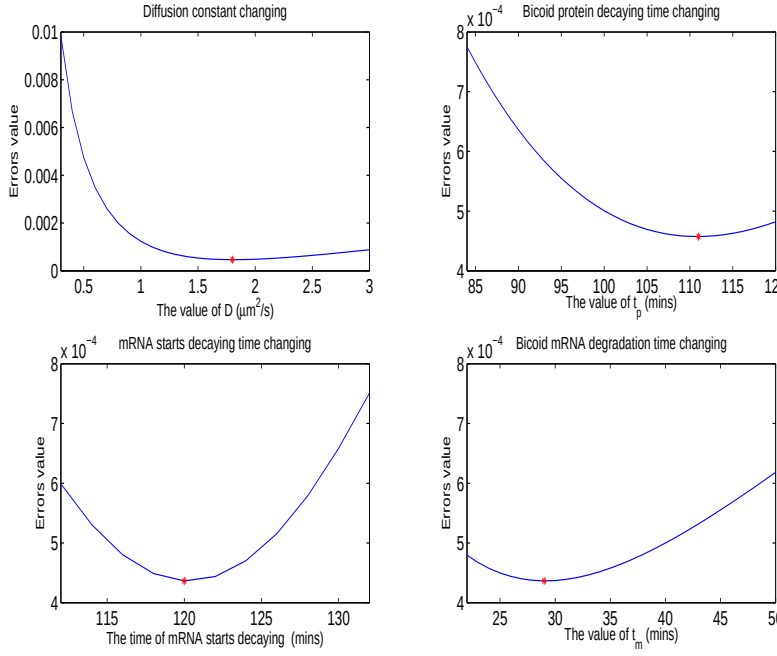


Fig. 5. Modelling errors as functions of the four parameters: diffusion constant(D), bicoid protein half-life (τ_p), the time of maternal mRNA decay onset (t_0) and maternal bicoid mRNA half-life (τ_m). The errors are computed during the eight stages of developmental cycle 14A, holding three of the four parameters at their best estimates from literature and varying the fourth.

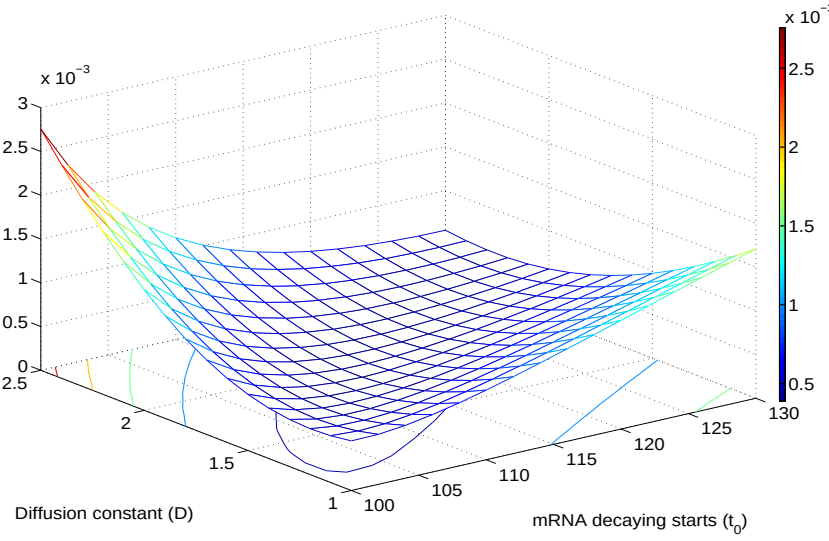


Fig. 6. Modelling errors in the joint space of diffusion constant and maternal mRNA decay onset time.

3.5 Finding Optimal Values for all the Parameters

In the previous sections, we discussed the parameter estimation by holding three of the four parameters at their best estimates from literature and changing the fourth. We also searched for the best combination of parameter values simultaneously on a regular grid. This search resulted in values closed to the results obtained previously: the diffusion constant $D = 1.83\mu m^2/s$, mRNA starts to decay $t_0 = 118$ mins, mRNA half-life $\tau_m = 28.4$ mins and bicoid protein half-life $\tau_p = 120$ mins.

4 Conclusion

Widely used models of how a profile of morphogen is established assume passive diffusion with a constant supply of morphogens at the source. The assumption of a constant source is unrealistic for a number of reasons. In this paper we have addressed a particular weakness of these models, *i.e.* decay of the source mRNA after a certain time, using the morphogen bicoid as an example. By matching the resulting model profile of bicoid to measured profile taken from [8], we show early results on how parameters of the diffusion model can be calculated. In the present study we have used single measurements of profiles from the FlyEx

database, which do not have uncertainties of the measurements quantified. The next step in this work would be to acquire uncertainties in bicoid profile measurements, arising from a distribution across a population of embryos, formulate the estimation problem in a probabilistic setting, and carry out posterior inference along the lines in [11].

References

1. Turing, A.: The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society B **237**(641) (1952) 37–72
2. Gurdon, J., Bourillot, P.: Morphogen gradient interpretation. Nature **413** (2001) 797–803
3. Yucel, G., Small, S.: Morphogens: Precise outputs from a variable gradient. Current Biology **16**(1) (2005) 29–31
4. Bergmann, S., Sandler, O., Sberro, H., Shnider, S., Schejter, E., Shilo, B., Barkai, N.: Pre-steady-state decoding of the bicoid morphogen gradient. PloS Biology **5**(2) (2007) 965–991
5. Hecht, I., Rappel, W., Levine, H.: Determining the scale of the bicoid morphogen gradient. PNAS **106**(6) (2009) 1710–1715
6. Gregor, T., Tank, D., Wieschaus, E., Bialek, W.: Probing the limits to positional information. Cell **130** (2007) 153–164
7. Surdej, P., Jacobs-Lorena, M.: Developmental regulation of bicoid mrna stability is mediated by the first 43 nucleotides of the 3' untranslated region. Molecular and Cellular Biology **18**(5) (1998) 2892–2900
8. Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M., Reinitz, J.: A database for management of gene expression data in situ. Bioinformatics (2004)
9. Holloway, D., Harrison, L., Kosman, D., Vanario-Alonso, C., Spirov, A.: Analysis of pattern precision shows that drosophila segmentation develops substantial independence from gradients of maternal gene products. NIH Public Access Author Manuscript (2006)
10. Lucchetta, E., Lee, J., Fu, L., Patel, N., Ismagilov, R.: Dynamics of drosophila embryonic patterning network perturbed in space and time using microfluidics. Nature **434** (2005) 1134–1137
11. Lawrence, N., Sanguinetti, G., Rattray, M.: Modelling transcriptional regulation using gaussian processes. NIPS **20** (2007) 785–792

On utility of gene set signatures in gene expression-based cancer class prediction

Minca Mramor[1], Marko Toplak[1], Gregor Leban[1], Tomaž Curk[1], Janez Demšar[1], Blaž Zupan[1,2]

[1] Faculty of Computer and Information Sciences, University of Ljubljana, Slovenia
[2] Dept. of Human and Mol. Genetics, Baylor College of Medicine, Houston, USA

Abstract. Machine learning methods that can use additional knowledge in their inference process are central to the development of integrative bioinformatics. Inclusion of background knowledge improves robustness, predictive accuracy and interpretability. Recently, a set of such techniques has been proposed that use information on gene sets for supervised data mining of class-labeled microarray data sets. We here present a new gene set-based supervised learning approach named SetSig and systematically investigate the predictive accuracy of this and other gene set approaches compared to the standard inference model where only gene expression information is used. Our results indicate that SetSig outperforms other gene set approaches, but contrary to earlier reports, transformation of gene expression data to the space of gene set signatures does not result in increased accuracy of predictive models when compared to those trained directly from original (not transformed) data.

1 Introduction

Methods to incorporate additional *domain knowledge* in the model inference process have from its early ages been central to machine learning research. Also referred to as *background knowledge*, its inclusion should increase model stability, predictive accuracy and interpretability.

In systems biology the sources of domain knowledge abound. They include information on gene structure and annotation, protein interactions, tissue localization, biological pathways, literature references, and other. From the onset of high-throughput data acquisition, bioinformatics has striven to include such additional knowledge in the discovery process. Consider, for instance, genome-wide gene expression analysis. From the first reports on utility of computational techniques such as clustering, the relevance of results was confirmed using function annotations [1]. Later, the procedure was formalized in *enrichment analysis*, where knowledge on groups of related genes, called *gene sets*, was used to identify groups including either over or under-expressed genes under specific experimental conditions [2]. Reporting enriched gene sets, rather than a list of differentially expressed genes, should yield stability, improve robustness across data sets of the same kind coming from different sources (labs), and help us in

gaining a deeper understanding of the underlying processes due to identification of affected pathways [3].

Gene set enrichment is by definition an explorative data analysis technique. If the task in genome-wide microarray analysis is class prediction, such as tumor classification, diagnosis and prognosis, standard supervised machine learning techniques should be used instead [4]. Early efforts in this domain directly applied machine learning to class-labeled expression data [5] and used gene expressions as features. Recently, a number of techniques have been proposed to incorporate the knowledge on gene sets in the model inference process, where each individual observation (*e.g.* tissue sample) should be described by features (*signatures*) that correspond to gene sets. These are computed from expression of its constituents (genes) and are then used for model inference. At present, these approaches can be classified based on whether they use class information when computing the signatures. Approaches that do not use class information include methods that compute average gene set expression [6], use principal component analysis (PCA) [7] or singular value decomposition [8, 9], while domain-enhanced analysis with partial least squares [7], PCA with relevant gene selection [10], activity scores based on condition-responsive genes [11], averages of expression values of genes supporting the gene set score [12] and ASSESS [13] do.

Similarly to gains in enrichment analysis, gene sets-based inference of predictive models should improve the stability and predictive accuracy. Interestingly, however, this has not yet been systematically tested across larger collections of data sets and across different methods. Also, there is a lack of a thorough comparison of such approaches with standard machine learning from the entire set of genes.

In the paper, we demonstrate the stages of development of a gene set-based supervised learning approach in crafting our own one (SetSig), and then report on systematic investigation to determine if this and five other knowledge-based techniques produce more accurate predictive models. Our test-bed incorporates 30 publicly available data sets, and uses standard evaluation and modelling procedures from supervised data mining. The results of our analysis are quite surprising and contradict initial reports on the superiority in accuracy of gene set-based predictive modelling [11–13].

2 Methods

2.1 Data sets

The study considered 30 cancer gene expression data sets from the Gene Expression Omnibus (GEO) [14]. All data sets have two diagnostic classes and include at least 20 samples, where each class was represented by at least 8 data instances. On average, the data sets include 44 instances (s.d.= 29.6). The GDS data sets with following ID numbers were used: 806, 971, 1059, 1062, 1209, 1210, 1220, 1221, 1282, 1329, 1375, 1390, 1562, 1618, 1650, 1667, 1714, 1887, 2113, 2201, 2250, 232, 2415, 2489, 2520, 2609, 2735, 2771, 2785 and 2842.

All data sets were preprocessed in the same manner. First, the probes measuring the expression of the same gene were joined and the average value of the expression over all probes was used. Second, in all data sets the gene expression values for each gene were normalized to zero mean ($\mu = 0$) and unit variance ($\sigma^2 = 1$).

2.2 Gene sets

We used the gene sets from the Molecular signatures data base (MSigDB v2.5) [2]. MSigDB includes five collections of gene sets that differ in the prior knowledge or the computational method used for creating them. We have considered collections C2 and C5, where gene sets where composed based on prior biological knowledge. From these we selected gene sets that include at least five genes for which the gene expression information was provided in the explored data set. Also, large (and possibly non-specific) gene sets that included more than 200 such genes were excluded from the analysis. As a result of this filtering, we used the following gene sets:

- C2cp: 639 gene sets belonging to canonical pathways (C2 collection). These gene sets are compiled by domain experts from the pathway data bases and are usually canonical representations of a biological process.
- C2C5: gene sets from the biological process and molecular function part of gene ontology (C5 collection) in addition to gene sets from C2cp. Depending on the number of genes in the specific data set, approximately 1.600 gene sets covering up to 7.900 genes met these criteria.

2.3 SetSig: sample characterization by gene set signatures

We here describe SetSig, a new approach to summarizing gene expression data into features based on gene sets. Our primary motivation was to construct a relatively simple method that does not rely on linear transformations and on search for gene groups within gene subsets which can potentially lead to overfitting.

Gene expression data consists of a number of samples S described by gene expressions, $f_S(g)$ (where g represents a gene) and the class value. SetSig transforms the data so that samples are described by gene set signatures, $f_S(G)$ (where G is a gene set) computed from the original gene expressions. The procedure for computation of $f_S(G)$ for a particular sample S and gene set G goes as follows:

1. Let C_1 and C_2 be sets of samples belonging to the first and to the second class, respectively.
2. Calculate the Pearson correlation coefficient between the expressions of genes from gene set G in the sample S and every sample from C_1 and from C_2. For a given gene set G, let R_1 and R_2 then be the corresponding sets of correlation coefficients, that is

$$R_1 = \{r_G(S, C) : C \in C_1\}, \qquad R_2 = \{r_G(S, C) : C \in C_2\},$$

where $r_G(S, C)$ is the correlation between $f_S(g_i)$ and $f_C(g_i)$ for $g_i \in G$.

3. The genes set G's signature for sample S, $f_S(G)$, is then computed as the Student's t-statistics for difference between R_1 and R_2:

$$f_S(G) = \frac{\overline{R_1} - \overline{R_2}}{\sqrt{s^2_{R_1}/N_1 + s^2_{R_2}/N_2}},$$

where N_1 and N_2 are the number of samples in C_1 and C_2, respectively.

Intuitively, coefficients in R_1 are high (low) if expressions of genes from gene set G in the sample S are similar to (different from) expressions of these genes in the samples from the first class. Coefficients in R_2 describe the similarities (differences) for the second class. Student's t-test measures whether the coefficients in R_1 differ from those in R_2 that is, how important are the genes from G for distinguishing between the two classes. The sign of the t-statistic is positive (negative) if the particular sample's gene expressions are more similar to those of first (second) class.

This procedure is used on each sample and for each gene set. The result is a set of samples described with gene set-based features, instead of by gene expressions. Notice that the same procedure is used for gene set signature construction for the training and testing set, where the signatures of testing instances are obtained using sets R_1 and R_2 computed on the training data. While SetSig directly addresses the data with binary class variable, it can be simply extended to multi-class prediction problems by construction of a separate classifier for each of the sample labels. In the paper we concentrate on the performance of the core method only and study only binary classification problems.

2.4 Other gene set signature transformation methods

In experiments we compared SetSig to other, previously published methods that use transformation of gene expression data sets to data sets comprising gene set scores. These transformations include:

1. Mean and Median [6], where each gene set is characterized with mean (median, respectively) expression of genes from the gene set.
2. ASSESS [13], which ranks sample's genes according to the differential probabilities of the two classes and scores gene sets as the deviation of a random walk from zero. The parametric model [13] was used for estimation of differential probabilities.
3. The first principal component of PCA [7] of genes in the gene set.
4. Activity scores based on condition-responsive genes (CORGs) [11], which differs from other evaluated methods by using only a subset of up or down-regulated genes from the gene set.

2.5 Estimation of predictive accuracy, classification, evaluation of results

Different supervised learning methods have been used to build class prediction models in the space of gene set signatures and in the space of gene expressions.

Models were built with support vector machines (SVMs) with linear kernel, a naive Bayesian classifier, a k-nearest neighbor learner, and a logistic regression learner. We report on the results for the SVM and logistic regression models, which outperformed models built with other supervised learning approaches. The results of other tested class prediction methods show similar trends.

We used leave-one-out validation to estimate the area under ROC curve (AUC) of the tested models. The same evaluation procedure was used across the entire set of 30 data sets. For each data set, the various methods were ranked. Statistical significances of differences between average ranks of tested methods were evaluated with the Nemenyi test and were visualized with critical distance graphs [15].

All supervised learning approaches were used as embedded in Orange data mining environment [16]. Orange was also used to implement SetSig and re-implement all other gene set-based supervised learning procedures investigated in this report.

3 Results

We first compared the predictive accuracy of class prediction models using SetSig transformed data sets with the C2cp and C2C5 gene set subsets with predictive accuracy of the models built with original gene expression data. For the latter, no feature selection or any additional data transformation was used. Figure 1 shows that SVM models built with original data sets perform significantly better than SetSig on the C2cp subset and better (but not significantly) for the C2C5 subset. As expected, SetSig performs better with larger number of gene sets (more biological knowledge), albeit the difference was not significant.

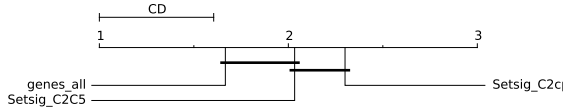


Fig. 1. Critical distance graph showing the average AUC ranks of SVM models on original gene expression data sets(genes_all) and data sets transformed by SetSig (either with gene sets C2cp or C2C5). Methods connected with bold lines are not significantly different ($\alpha = 0.05$).

Figures 2 and 3 include the results for all gene set-based transformations listed in Sec. 2.4 for the SVM and logistic regression models, respectively. Gene sets in C2C5 were used as the models built with them performed better in the experiments with SetSig reported above. Nemenyi test identifies two groups of insignificantly different methods connected with a bold line in Figure 2. Inference from gene expression without gene set transformation performs best, although

the difference is only significant for two of the six gene set-based methods (PCA and ASSESS). The difference between all gene set methods is statistically insignificant. Of all the tested methods, SetSig performed best. Similar trends can be observed in Figure 3 for the models built with logistic regression. Again, models built with the original gene expression data preform best. The difference in the average ranks is significant for two of the gene set transformation methods (Median and CORGs). SetSig outperforms other gene set transformation methods and is significantly better than the Median approach.

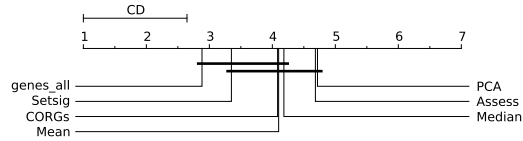


Fig. 2. The average AUC ranks of SVM models on original gene expression data sets (genes_all) and transformed using a variety of gene set-based transformation methods.

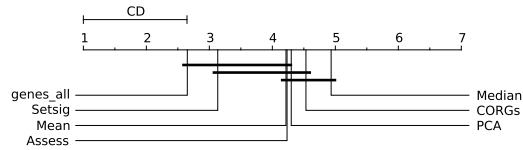


Fig. 3. The average AUC ranks of logistic regression models on original gene expression data sets (genes_all) and transformed using a variety of gene set-based transformation methods.

Gene set-based approaches use only a subset of genes from the original expression data sets. One reason for poorer performance of these approaches could have been that some informative genes are left out. We tested this by evaluating the accuracy of predictive models built directly from gene expressions but using only a subset of genes. We have examined the following subsets in this way: (1) genes present in C2cp (genes_C2cp), (2) genes not present in C2cp (genes_notC2cp), (3) genes present in C2C5 (genes_C2C5), (4) genes not present in C2C5 (genes_notC2C5), and (5) all genes (genes_all).

The average ranks of the models built with the above mentioned subsets and the differences between them are shown in Figure 4. The average ranks of AUC of models built with different subsets of genes are very similar. No statistically significant differences were detected.

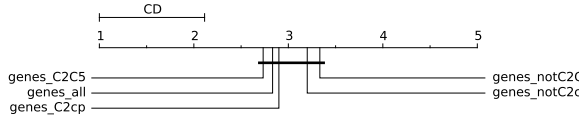


Fig. 4. The average ranks of AUC-scored classifiers that use different subsets of genes. The differences are statistically insignificant.

4 Discussion

Our experimental results indicate that transformation of gene expression data to the space of gene set signatures does not result in increased accuracy of predictive models when compared to those trained from original (not transformed) data. In fact, the latter, "gene-set free" approach consistently ranked higher in our experiments. Of all the tested gene set approaches, SetSig's performance was closest to that of using all genes.

These results come as a surprise. First, in explorative data analysis, the utility of gene sets is motivated by gains in interpretability, and also by gains in stability and robustness of results, even when compared across data sets obtained from different laboratories [17].

Next, several recently published papers explicitly report that their gene set approaches over-perform the gene-centric approach. Closer inspection shows that these assertions are not a result of systematic study, and either used a very limited number of data sets in the study [12, 13, 11], or, as in the most recent report, are based on too restrictive gene selection (feature set selection of only a handful genes in gene-centric approach) prior to learning [11]. But even with such lack of systematic testing, all the present evidence reported votes in favor of gene set-based approaches.

Finally, we would in general (albeit naively) expect to gain with any inclusion of additional (background) knowledge in machine learning. However, in frameworks described in this paper such knowledge is used to transform, rather than complement the problem domain. We can think of a number of other reasons why the utility of gene sets with respect to predictive accuracy fails:

1. Gene sets do not include some highly class-informative genes.
2. There are too many gene sets.
3. Some gene sets are very similar to each other.
4. Gene set signature construction methods lose information.
5. Number of samples (instances) is too low to reliably estimate gene set scores.
6. Biological knowledge of the genes is incomplete. Gene sets and pathways used are not specific enough to represent biological processes that distinguish between different cancer types.

We can reject reason (1) based on results on gene-centric approach that used genes from different sets (Figure 4), where no significant differences were observed. Facts stated in (2) and (3) can hurt supervised learning, but gene-centric

approaches must deal with the same kind of problems (abundance of genes, many of which are co-expressed genes). Due to (4) we have tested six different approaches, including very promising and elaborate ones such as CORGs. (5) clearly deserves further investigation. Previous studies have already shown that supervised learning methods may fail due to low sample size [18, 19]. Finally (6), despite incompleteness of biological knowledge on genes, we would expect that additional information in the form of gene sets should help us in inference of reliable classifiers, even more for the methods like CORGs which remove genes that do not contribute to class differentiation from the gene sets.

5 Conclusion

The reasons why gene set-based transformations for supervised learning from gene expression data sets fail when compared to gene-centric learning seem elusive. In fact, they do not fail, but rather – contrary to our expectations and to several recent reports – do not surpass the more standard and direct learning from gene expression profiles. Yet, predictive performance is not the only issue here, and gene set-based predictive models can significantly gain with regard to ease of interpretation and information they provide to biologists and clinicians. We have indeed observed that just like for gene-centric models [20] we could construct very simple and highly-predictive visual models using only a few gene set signatures. We can thus conclude that knowledge on gene sets may be a useful resource for supervised microarray data analysis, but that methods for its inclusion in model inference require further studying and improvements, specifically in terms of gains in predictive accuracy.

Acknowledgements

This study was funded by the program and project grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112).

References

1. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. PNAS **95**(25) (1998) 14863–8
2. Subramanian, A., Tamayo, P., Mootha, V.K., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS **102**(43) (2005) 15545–50
3. Nam, D., Kim, S.Y.: Gene-set approach for expression pattern analysis. Brief Bioinform **9**(3) (2008) 189–97
4. Simon, R., Radmacher, M.D., Dobbin, K., McShane, L.M.: Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. J Natl Cancer Inst **95**(1) (2003) 14–8
5. Brown, M., Grundy, W.N., Lin, D., et al.: Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS **97**(1) (2000) 262–7

6. Guo, Z., Zhang, T., Li, X., et al.: Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinformatics **6** (2005) 58
7. Liu, J., Hughes-Oliver, J.M., Menius, J. A., J.: Domain-enhanced analysis of microarray data using GO annotations. Bioinformatics **23**(10) (2007) 1225–34
8. Tomfohr, J., Lu, J., Kepler, T.B.: Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics **6** (2005) 225
9. Bild, A.H., Yao, G., Chang, J.T., et al.: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature **439**(7074) (2006) 353–357
10. Chen, X., Wang, L., Smith, J.D., Zhang, B.: Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics **24**(21) (2008) 2474–81
11. Lee, E., Chuang, H.Y., Kim, J.W., et al.: Inferring pathway activity toward precise disease classification. PLoS Comput Biol **4**(11) (11 2008) e1000217
12. Efron, B., Tibshirani, R.: On testing the significance of sets of genes. Ann Appl Stat **1**(1) (2007) 107–29
13. Edelman, E., Porrello, A., Guinney, J., et al.: Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. Bioinformatics **22**(14) (2006) e108–16
14. Barrett, T., Troup, D.B., Wilhite, S.E., et al.: NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucl. Acids Res. **35** (2007) 760–5
15. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of machine learning research **7**(jan) (2006) 1–30
16. Demšar, J., Zupan, B., Leban, G.: Orange: From experimental machine learning to interactive data mining, white paper (2004)
17. Manoli, T., Gretz, N., Grone, H.J., et al.: Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics **22**(20) (2006) 2500–6
18. Ein-Dor, L., Kela, I., Getz, G., et al.: Outcome signature genes in breast cancer: is there a unique set? Bioinformatics **21**(2) (2005) 171–178
19. Hanczar, B., Dougherty, E.R.: Classification with reject option in gene expression data. Bioinformatics **24**(17) (2008) 1889–1895
20. Mramor, M., Leban, G., Demšar, J., Zupan, B.: Visualization-based cancer microarray data classification analysis. Bioinformatics **23**(16) (2007) 2147–2154

Accuracy-Rejection Curves (ARCs) for Comparison of Classification Methods with Reject Option

Malik-Sajjad-Ahmed Nadeem[1,2,5,⋆], Jean-Daniel Zucker[2,3,6], Blaise Hanczar[4]

(1)LIM&Bio, UFR de Sante, Medecine et Biologie Humaine (SMBH) - Leonard de Vinci, University Paris 13, 74 rue Marcel Cachin, 93017 Bobigny Cedex, FRANCE.
(2)INSERM U872 Equipe 7, Centre Recherches des Cordeliers, 15 rue de l'Ecole de Medecine, University Paris 6, 75005 Paris, FRANCE.
(3)UPMC University Paris 6, UMRS 872, NUTRIOMIQUE, CRC, 75006 Paris, FRANCE.
(4)LIPADE University Paris Descartes, 45 rue des Saint-Peres,75006 Paris, FRANCE.
(5)Department of Computer Sciences & IT, University of Azad Jammu & Kashmir Muzaffarabad, 13100 Muzaffarabad, Azad Jammu & Kashmir, PAKISTAN.
(6)Institut de Recherche pour le Developpement(IRD) IRD, UMI 209, UMMISCO, IRD France Nord, F-93143, Bondy, FRANCE.

1 Introduction

Microarray classification is a topic of great interest now-a-days in medical and bioinformatics research. Microarrays simultaneously measure the mRNA expression level of thousands of genes in a cell mixture at certain times and in different environmental conditions. One of the main characteristic of this kind of data is the huge disproportion between the number of examples (generally 10 to 100 microarrays by experiment) and number of features (several thousands of genes). Microarrays are used in many fields of medical research. Among the most prominent and useful applications is the prediction of a biological parameter based on the gene-expression profile. For example, by comparing the expression profiles of different tissue types we can predict different biological parameters like different types of tumors with different outcomes, survival time of a cancer patient after a therapy, weight loss prediction after a diet control and/or bariatric surgery etc. and hence assist in the selection of a therapeutic treatment [7, 2, 19].

A large number of methods, from machine learning, have been successfully applied to classify microarrays, Diagonal Linear Discriminant Analysis (DLDA) and k-nearest neighbors [7], Support Vector Machine [9], Random Forests [3] etc. The performances of these classifiers are measured by their accuracy to predict the true class. This accuracy is estimated by re-sampling procedures like cross-validation or bootstrap. A natural question is which one is the best classifier for microarray based classification? Unfortunately the answer is not easy. Several comparative studies have been published. Man et al. [17] claim that Support

⋆ Corresponding author. E-mail addresses: mnadeemsajjad@gmail.com, msajjad-nadeem@yahoo.com. I am a Ph.D candidate in the filed of biomedical informatics.

Vector Machine (SVM) and Partial Least Squares Discriminant Analysis (PLS-DA) have the best accuracy. Dudoit et al. [7] show that simple methods like DLDA and k-nearest neighbors produce good results, whereas Statnikov et al. [18] conclude the superiority of SVM. Probably the more confident conclusion is given by both Lee et al. [16] and Huang et al. [12] that there is no classifier uniformly better than the other. Moreover, all these studies are based on the error rates of obtained classifiers. But the error rate is not the only metric of interest to measure the quality of a classifier. Generally, in medical application, a reject option is added to the classifiers. When the prediction of an example is not safe, the classifier reject this example and does not assign any class to it. Reject option introduced by Chow [5] states that refrain from taking decision for samples whose decision is less confident in order to reduce error probabilities. Friedel et al. [8], Hanczar et al. [10] and others used reject option in their methods for improvement in prediction accuracy of classifiers and proved that reject option considerably enhances the prediction accuracy of classifiers. So, the performance of these classifiers depend on both accuracy and rejection rate. But we don't have a tool to compare the performances of different classifiers with reject option. In this paper we propose a methodology called ARC that compares the performance of classifiers based on their accuracies and rejection rates. According to our knowledge, there is no comparison study including classifiers with reject option in the literature. A general assumption is that the comparison of classifiers is the same with and without reject option. In this paper, we test this assumption and show that it is wrong. Our assumption is that rejection has different impact on the accuracy of different classifiers, and the best classifier depends also on the quantity of rejection. The ARC methodology proposes to compare the performances of classifiers in the function of their reject rate by considering different reject areas ranging from 0.2% to 100% reject rates. Our experimental results based on diverse pure artificial data and artificial data sets synthesized from real data show that the proposed comparison of different classifiers (with reject option) is advantageous for the selection of best available classifier for a given data.

2 Classification with Reject Option

Chow [4] introduces the concept of reject option. Consider a binary classification problem where each example falls in one of the categories. The performance of a classifier is measured by its error rate. The classifier minimizing the error is called the Bayes classifier.

If the accuracy of the Bayes classifier is not sufficient for the task at hand, then one can take the approach not to classify all examples, but only those for which the posterior probability is sufficiently high. Based on this principle, Chow [5] presented an optimal classifier with reject option. A rejection region is defined in the feature space and all examples belonging to this region are rejected by the classifier. The classifier rejects an example if the prediction is not sufficiently reliable and falls in the rejection region. There is a general relation between the

error and rejection rate: According to Chow [5] the error rate decreases monotonically while the rejection rate increases. Based on this relation, Chow proposes an optimal error versus reject tradeoff. In classifier with rejection option, the key parameters are the thresholds that define the reject areas. Landgrebe et al. [14] Dubuisson and Masson [6], Hanczar and Dougherty [10] and others proposed several strategies to find an optimal reject rule. In our work, we do not deal with the problem of optimal tradeoff between error and rejection. In our approach, we used different rejection areas and computed resulting accuracies. We varied the size of rejection area from 0% to 100% by an increment of 0.2% resulting in 500 rejection areas. To represent the results we plotted the rejection windows against obtained accuracies.

3 Comparing Classifiers with Reject Option

The performances of classifiers are measured by their accuracy to predict the true class. Several studies (some of them mentioned in introduction) claim the superiority of different classifiers. All the comparitive studies are based on the error rates of obtained classifiers but error rate is not only the measure to judge a classifier's performance. The performance of a classifier depends heavily on the data too. So, for each classification task, a comparison study should be done to determine the best classifier. In case of classification with reject option, the accuracy also depends on the reject rate. In this paper we propose a classifiers' comparison method in the scenario of reject option. The idea is to watch the accuracies of the classifiers as the function of their reject rate. Based on this idea we define 3 different situations:

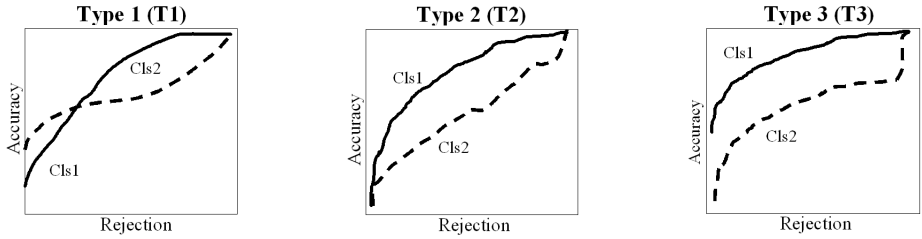


Fig. 1. Illustration of the 3 cases of possible Accuracy-Rejection Curves (ARCs).

1. Case 1: A classifier (say Cls_1) initially performs worse than another classifier (say Cls_2). By opting reject option, Cls_1 outperforms Cls_2. Name this crossing over as $T1$ type Accuracy-Rejection Curve(ARC).
2. Case 2: Without selecting to reject or rejecting to some extent both the classifiers Cls_1 and Cls_2 perform approximatley same but with more and more rejection, one of the classifier increases its performace more rapidly than other. Call this diversion as $T2$ type ARC.

3. Case 3: If Cls_1 and Cls_2 are very much distinct in their performance without rejection but the reject option does affect identically to both of them. Name these curves as $T3$ type ARCs.

In this paper we propose that the performances of classifiers can also be represented by 2-dimensional Accuracy-Rejection Curves (ARCs) where the axes are their accuracies and rejection rates. Figure 1 illustrates the 3 different cases that we defined. To select the best available classifier using ARCs for a problem in hand, a measure (desired accuracy, acceptable rejection rate) should be known. If desired accuracy is known we move horizontally on ARCs plot and select the available classifier with least rejection rate. We select the classifier with maximum prediction accuracy for a given rejection rate.

3.1 Data

Our experiments are based on two kinds of data: pure artificial data generated using Gaussian models, synthetic data generated using parameters estimated from real microarray data using Expectation-Maximization (EM) algorithm from microarray studies: colon cancer data (Alon et al.), lymphoid malignancy (Shipp et al.) and acute myeloid leukemia" (AML) and acute lymphoblastic leukemia (ALL) (Golub et al.). Details of artificial data generation proceudre and different settings and description of the parameters can be found on the companion website `http://bioinfo.nutriomics.org/~sajjad/ARC/`.

The experiments on real data require the use of sampling methods to estimate the error rate and it has been show that these methods are inaccurate for small-sample problems rather the use of synthetic data give more accurate error estimation [11]. So we don't use real microarray data in our experiments.

For each classification problem, we generate data with 20 features, called noise free features. In real microarrays most of the genes are irrelevant for the classification task in hand [1, 9, 22, 15]. So to have a more realistic aspect, 380 irrelevant or noise features $d_{irrF} = 380$ are added to artificial datasets. A noise feature follows the same Gaussian distribution for the two classes $N(\mu; \sigma)$. The generated data contain N examples, 400 features where 380 are noise features and 20 are noise free features.

3.2 Experimental Design:

We used following decorum in our experimental design.

1. Generate class-labelled train data n_{tr}containing 50, 100 or 200 examples and a total of $D = D_{nf} + D_n$ features.
2. Generate test data n_{ts}containing 10000 examples and a total of $D = D_{nf} + D_n$ features.
3. Find 20 or 40 best features by using t-test feature selection method on DT_r and reduce train data by selecting only $d_{sel} = 20$ best features from train data set.

4. Reduce test data by driving the same best features from test dataset DT_s.
5. Apply a classification rule to build a classifier Cls from DT_r according to most widely used classification rules for microarray analysis including Support Vector Machine Linear kernel (SVM-Linear); Support Vector Machine Radial kernel (SVM-Radial); Linear Discriminant Analysis (LDA); Quadratic Discriminant Analysis (QDA); random Forest (RF).
6. Compute true error rate/rejection rates of the underlying model on test data.
7. Repeat step 6 for all sizes of rejection windows $R_{win} = \{0.002, 0.004, 0.006, 0.008, ...1.000\}$
8. All steps 1-7 iterated 100 times.
9. Final result is averaged from all iterations.

We randomly generated 100 different data sets in each case and then these 100 replications are used for classification using classification rule (SVM-Linear, SVM-Radial, LDA, QDA, etc).

4 Results and Discussion

The experiments use both pure synthetic data and synthetic data based on real microarray patient data. The experiments on synthetic data permit very accurate estimations of the error and rejection rates.

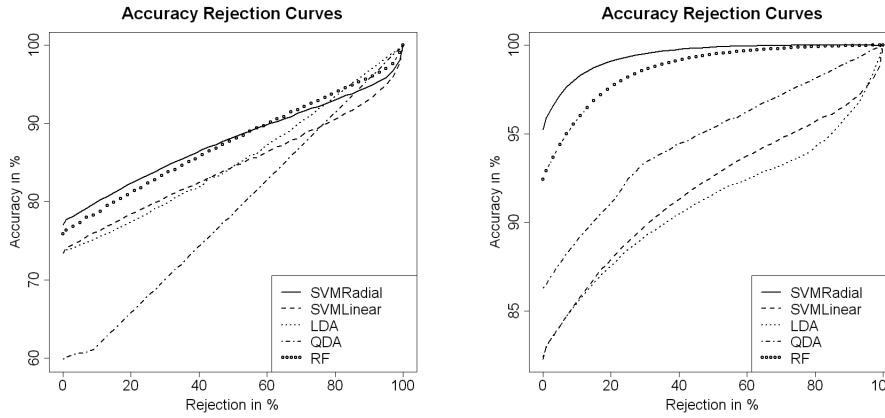


Fig. 2. A(left):Rejection verses Accuracy curve on linear, non-correlated data with 1 Gaussian per class where train set = 50 examples and test set= 10000 examples. B(right):Rejection verses Accuracy curve on non-linear, correlated data with 1 Gaussian per class where train dataset =100 examples and test dataset = 10000 examples.

In each of the following figures we plot average rejection versus average accuracy for all classification rules $R = 5$ and for one of the data sets. Here we

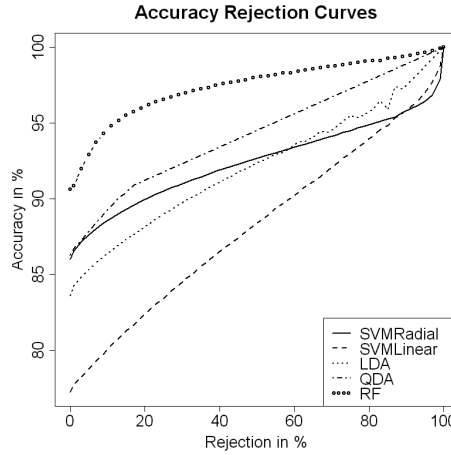


Fig. 3. Rejection verses Accuracy curve on Synthetic data from colon cancer patient dataset with 5 Gaussians per class where train dataset = 200 examples and test dataset= 10000 examples.

present some typical results while leaving the complete results on the companion website `http://bioinfo.nutriomics.org/~sajjad/ARC/`. In the plots solid lines represent the accuracy rejection curve of SVM with Radial kernel, dashed lines show SVM with Linear kernel, dotted lines are of LDA, dashed-dotted lines are of QDA, and filled-circle lines represent RF. In Figure:2A we notice that SVM-Radial without rejection (0% rejection) produces around 87% accuracy and RF without rejection (0% rejection) results 85% accuracy. By opting to reject around 50% RF becomes better classifier than SVM-Radial. Also an interesting point here in Figure:2A is that with 45% rejection rate both LDA and SVM-Linear behave similarly as for as accuracy is concerned. But after 45% rejection, LDA outperforms SVM-Linear. Figure:2A depicts that initially without rejection LDA and SVM-Linear have almost identical accuracies. While rejecting on 3% and more samples SVM-Linear performs better than LDA. Figure:2B shows that LDA and SVM-Linear produce similar accuracies starting from without rejection (0% rejection) to 18% rejection but from 19% rejection SVM-Linear starts performing much better than LDA. In Figure:3, while comparing LDA and SVM-Radial, we found the situation where curves of LDA and SVM-Radial cut each other making LDA better than SVM-Radial. Also this figure does show that on evaluating the performances of QDA and SVM-Radial, QDA outperforms SVM-Radial on having reject option. Each of our result reflects that as we reject more, we get more and more accuracy. Not all the classification rules used here respond identically to reject option. By analysing three presented figures in this section, we have interesting results where different classification rules respond differently different at different rejection rates. In our study some re-

spond more quickly and we get more accurate classification than that of others. Empirical results show that most of the times one or more classification rules outperform the other(s). On the basis of above presented results and discussion we can have three types of ARCs proposed in the section Comparing Classifiers with Reject Option. The identification of these types of curves is advantageous in several ways. First: during the selection of suitable classifier for a classification problem if $T1$ curves are available then the classifier which outperforms the others should be given priority. Second: In case of $T2$ curves, we may reject upto desired limit and then the classifier with high performance may be utilized. Third: When $T3$ curves are there then at a given rejection extent, the classifier with higher performance should be selected for that specific dataset for which the comparison was made.

In Tables on companion website `http://bioinfo.nutriomics.org/~sajjad/ARC/` we summarize our all the 90 experiments based on the above mentioned categories of curves. While experimenting with pure artificial data we noticed that in 72 experiments we have 40 times the situation when one or more classifier outperforms the other (by crossing over of curves i.e. category $T1$). Also an interesting point is that we have 59 situations where without or with some rejection, two or more classifiers perform almost identically. But with more or less rejection, one of the classifier improves its prediction capability more promptly than the other (category $T2$). Here we have only 12 cases where $T3$ type curves are present in the results. In total of 90 experiments we found 43 times when one or more classifier outperforms the other through $T1$ curves. We also experienced 64 $T2$ curves. The presence of more $T2$ and $T3$ curves reflects that the use of reject option in comparison of classifiers is extremely fruitful and in most of the cases aids in more optimal classifier selection. The presence of 22 $T3$ shows that sometimes rejection affects almost identicaly on the performances of the classifiers and there remains no significant change in the performances of two classifiers as compared to each other.

5 Conclusion

In this study we introduce the accuracy - rejection curves (ARCs) that allow to accurately represent the performance of classifiers. We see that it is necessary to watch both accuracy and rejection rate to compare two classifiers. On the basis of our empirical results we categorize the classifiers comparison into three types ($T1$, $T2$, $T3$ types ARCs). We made classifiers comparisons on a high number of experiment based on artificial data for 500 different reject areas ranging from 0.2% to 100% reject rates. We use different settings of parameters for pure synthetic data to construct different kinds of classification problems (linear and non-linear, correlated and non-correlated features with train sets). For synthetic data from real patients' data, model's parameters depend on the real data. In our results the presense of large number of T1 and T2 types of ARCs shows that ARCs are of interest while comparing classifiers' performances. Small number of T3 type ARCs reflects that there are some possibilities of no significinat change

in performance of a classifier while using reject option but the chances remain very little. While comparing the classifiers' performances, the extent upto which one can allow the rejection is still to be addressed. Also, obtaining optimal reject area is still an open question and needs further exploration. In function of the rejection rate, the conclusion of the comparison can be different.

References

1. 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.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 286, 531–537 (1999)
2. Braga-Neto, U.M., Dougherty, E.R.: Is cross-validation valid for small-sample microarray classification? Bioinformatics. 20(3), 374–380 (2004)
3. Breiman, L.: Random Forests. Machine Learning. 45, 5–32 (2001)
4. Chow, C.K.: An Optimum Character Recognition System using Decision Functions. IRE Trans. on Electronic Computers. EC-6, 247–254 (1957)
5. Chow, C.K.: On Optimum Error and Reject trade-off. IEEE Trans. on Information Theory. IT-16(1), 41–46 (1970)
6. Dubuisson, B., Masson, M.: A Statistical Decision Rule with incomplete Knowledge about Classes. Pattern Recognition. 26(1), 155–165 (1993)
7. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. Journal of the American Statistical Association. 97, 77–87 (2002)
8. Friedel, C.C., Ruckert, U., Kramer, S.: Cost Curves for Abstaining Classifiers. In Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning (2006)
9. Furey, T.S., Christianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support Vector Machine Classification and validation of Cancer Tissue Samples using Microarray Expression Data. Bioinformatics. 16(10), 906–914 (2000)
10. Hanczar, B., Dougherty, E.R.: Classification with Reject Option in Gene Expression Data. Bioinformatics. 24 no. 17, 1889–1895 (2008)
11. Hanczar, B., Hua, J., Dougherty, E.R.: Decorrelation of the True and Estimated Classifier Errors in High-Dimensional Settings. EURASIP Journal on Bioinformatics and Systems Biology. 2007, 12 pages (2007)
12. Huang, X., Pan, W., Grindle, S., Han, X., Chen, Y., Park, S.J., Miller, L.W., Hall, J.: A comparative study of Discriminating Human Heart Failure Etiology using Gene Expression profiles. BMC Bioinformatics. 6, 205 (2005)
13. Isaksson, A., Wallman, M., Gransson, H., Gustafsson, M.G.: Cross-validation and Bootstrapping are unreliable in small Sample Classification. Pattern Recognition Letters. 29(14), 1960–1965 (2008)
14. Landgrebe, T.C.W., Tax, D.M.J., Paclk, P., Duin, R.P.W.: The interaction between Classification and Reject Performance for Distance-based Reject-option Classifiers. Pattern Recognition Letters Pages. 27(8), 908–917 (2006)
15. Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G.: Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics. 17, 11318–1142 (2001)
16. Lee, J.W., Lee, J.B., Park, M., Songa, S.H.: An extensive Comparison of recent Classification tools applied to Microarray Data. Computational Statistics & Data Analysis. 48, 869–885 (2005)

17. Man, M.Z., Dyson, G., Johnson, K., Liao, B.: Evaluating Methods for Classifying Expression Data. Journal of Biopharmaceutical Statistics. 14, 1065–1084 (2004)
18. Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S.: A comprehensive evaluation of Multicategory Classification methods for Microarray Gene Expression Cancer diagnosis. Bioinformatics. 21, 631–643 (2005)
19. Wang, L., Chu, F., Xie, W.: Accurate Cancer Classification Using Expression of Very Few Genes. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 4(1), 40–53 (2007)
20. Egan, J.P.: Signal detection theory and ROC analysis, Series in Cognition and Perception. Academic Press, New York. (1975)
21. Swets, J.A., Dawes, R.M., Monahan, J.: Better decisions through science. Scientific American. 283, 82-87 (2000)
22. Zhou, X., Mao, K.Z.: LS Bound based gene selection for DNA microarray data. Bioinformatics. 21(8), 1559-1564 (2005)

Predicting the functions of proteins in PPI networks from global information

Hossein Rahmani[1], Hendrik Blockeel[1,2], and Andreas Bender[3]

[1] Leiden Institute of Advanced Computer Science, Universiteit Leiden,
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
hrahmani@liacs.nl,blockeel@liacs.nl
[2] Department of Computer Science, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, 3001 Leuven, Belgium
[3] Division of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research,
Leiden University
2333 CC Leiden The Netherlands
andreas.bender@pharma-it.net

Abstract. In this work we present a novel approach to predict the function of proteins in protein-protein interaction networks. We classify existing approaches into inductive and transductive approaches, and into local and global approaches. As of yet, among the group of inductive approaches, only local ones have been proposed. We here introduce a protein description formalism that also includes global information, namely information that locates a protein relative to specific other proteins in the network. The method is benchmarked on four datasets and we found that on these datasets classification according to precision and AUC values indeed improves over the benchmark methods employed.

1 Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g. to predict which proteins are suitable drug targets, according to an analysis of the resulting network. One particular machine learning task that has been considered is predicting the functions of proteins in the network.

A variety of methods have been proposed for predicting the classes of proteins. On a high level we can distinguish two types of approaches, namely inductive and transductive ones. Inductive learning approaches, also called model-based approaches, construct a model (a mathematical function) that maps a description of a protein onto its functions. Transductive approaches, on the other hand, immediately make predictions for the proteins in the network, without going through the intermediate stage of constructing a model that can be used afterwards for making predictions. The difference between these two will be described more formally in the next section.

Transductive approaches are often "global": information on the whole network is taken into account when making predictions. The inductive approaches that have been used up till now are typically local, in the sense that the description of a protein (from which its labels are to be predicted) contains information about the local neighborhood of the protein, not about the network as a whole. This is not an inherent property of inductive approaches, though; one might just as well try to construct a description that contains global information. Accordingly, in this paper we explore the usefulness of one particular kind of global information for the task of protein function prediction.

The paper is structured as follows. In Section 2 we define the learning problem formally. In Section 3 we briefly review approaches that have been proposed before to solve this problem. In Section 4 we present a new inductive learning approach; we do not present any new learning algorithms but a new description format of proteins, which contains global rather than local information. In Section 5 we empirically evaluate the performance of several learning algorithms when using this format, and, as a control experiment, compare this performance to that of a previously proposed approach. We present our conclusions in Section 6.

2 Problem Statement

Mathematically, PPI networks are graphs, and the problem we consider is that of predicting the labels of nodes in this graph.

Consider an undirected graph G with node set V and edge set E, where each node $v \in V$ is annotated with a description $d(v) \in D$ and, optionally, a label $l(v) \in L$. We assume that there exists a "true" labelling function λ from which l is a sample, that is, $l(v) = \lambda(v)$ where $l(v)$ is defined.

In **transductive** learning, the task is to predict the label of all the nodes. That is, given the graph $G = (V, E, d, l)$, with l a partial function, the task is to construct a completed version $G' = (V, E, d, l')$ with l' a complete function that is consistent with l where $l(v)$ is defined.

In practice, there is an additional constraint that l' should approximate λ. This is imposed by some optimization criterion o, the exact form of which expresses assumptions about λ. For instance, o could express that nodes that are directly connected to each other tend to have similar labels by stating that the number of $\{v_1, v)2\}$ edges where $l'(v_1) \neq l'(v_2)$ should be minimal. The assumptions made about λ are called the bias of the transductive learner.

In **inductive learning**, the task is to learn a function $f : D \to L$ that maps a node description $d(v)$ onto its label $l(v)$. That is, given $G = (V, E, d, l)$, we need to construct $f : D \to L$ such that $f(d(v)) = l(v)$ when $l(v)$ is defined, and f is defined for all elements of D. Note that f differs from l in that it maps D, not V, onto L. This implies, for instance, that it can also make predictions for a node v that was not in the original network, as long as $d(v)$ is known.

Besides the bias expressed by the optimization criterion o (which may still be present), there is now also a bias imposed by the choice of D: whenever two

different nodes have the same description, they are assumed to have the same labels: $d(v_1) = d(v_2) \Rightarrow \lambda(v_1) = \lambda(v_2)$. Additionally, the learning algorithm used to learn f has its own inductive bias [1]: given exactly the same inputs, two different learning algorithms may learn different functions f, according to assumptions they make about the likely shape of f.

Thus we have three types of bias. Transductive learners have a transductive bias, which is implied by the choice of the optimization criterion o. Inductive learners have a description bias, imposed by the choice of d, and an inductive bias, imposed by the choice of the learning algorithm that is used to learn f from $(d(v), l(v))$ couples. In this paper we will explore for one particular description function d wether it represents a suitable description bias.

In the context of protein function prediction in PPI networks, the nodes v are proteins; the descriptions $d(v)$ can be any description of the protein that can be derived from the network structure (no additional information, such as the protein structure, is assumed to be available; we assume we learn from the network structure only); the labels $l(v)$ are sets of protein functions.

Note that many proteins have more than one function; this is why a node label can be any set of functions. Most off-the-shelf machine learning techniques can only learn classifiers that predict a single value, not a set of values. The fact that node labels are sets may seem to form a problem in this respect. To remedy this situation, if we have n possible functions, the task of predicting a subset of these functions can easily be transformed into n single-function prediction tasks: for each possible function a binary classification task is then constructed where nodes are to be assigned the class true or false depending on whether the protein has that function or not. This is the setting we will focus on in this paper.

3 Related work

Among transductive approaches to the protein function prediction problem, the Majority Rule approach has a prominent role [2]. This method assigns to a protein those functions that occur most frequently among its neighbors (typically a fixed number of functions is predicted, for instance, the three most frequently occurring functions in the neighborhood). One problem with this approach is that it only considers neighbors of which the function is already known, ignoring all others. To address this problem, global optimization-based function prediction methods have been proposed. Any probable function assignment to the whole set of unclassified proteins is given a score, counting the number of interacting pairs of nodes with no common function; the function assignment with the lowest value will be the best assignment [3, 4].

Another improvement over the original implementation was made by observing higher-level interactions [5]. Level k interaction between two proteins mean that there is a path of length k between them in the network. Proteins that have both a direct interaction and shared level-2 interaction partners have turned out to be more similar to each other. Taking this further, one can make the assumption that in dense regions (subgraphs with many edges, relative to the

number of nodes) most nodes have similar functions. This has led to clustering approaches which first cluster the networks (with clusters corresponding to dense regions), and subsequently predict the function of unclassified proteins based on the cluster they belong to [6, 7].

Among the inductive approaches, Milenkovic et al.'s graphlet-based approach [8] has been used in the area of protein function predictions. The node description $d(v)$ that is built here, in their terminology the "signature vector", describes the local neighborhood of the node in terms of so-called graphlets, small graph structures as a part of which each node occurs. Most other inductive approaches use similar signatures. Typical for them is that they describe only the local structure of the network near the node to be predicted, however remote changes in the network do not influence the signature at all.

4 A global description of proteins

In this work we will now introduce an inductive approach that uses global node descriptions to the area of protein-protein interactions; that is, any change (e.g., addition or removal of an edge) in the network, wherever it occurs, may influence a node's description. Our hypothesis is that the inclusion of additional information will improve the function prediction of unknown nodes which will be investigated in the following in detail.

We describe a node as follows. Assume that there are n nodes in the network, identified through numbers 1 to n. Each node is then described by an n-dimensional vector. The i'th component in the vector of a node v gives the length of the shortest path in the graph between v and node i.

It has been hypothesized before that shortest-path distances are relevant in PPI network analysis; for instance, Rives and Galitski [9] cluster nodes based on shortest-path distance profiles. As of yet, however, such shortest-path distances have not been considered in the context of inductive learning of protein function predictors which is the reason for the current work.

A potential disadvantage of this method is that in large graphs, one gets very high-dimensional descriptions, and not all learners handle learning from high-dimensional spaces well. It is possible, however, to reduce the dimensionality of the vector by only retaining the shortest-path distance to a few "important" nodes. This essentially represents a feature selection problem. A node i is important if the shortest-path distance of some node v to i is likely to be relevant for v's classification. If the feature f_i denotes the shortest path distance to node i, one possible measure of the relevance of f_i for the label of a node (which is a set of functions) is the following.

For each function j, let G_j be the set of all proteins that have that function j. Let $\bar{f}_{ij}$ be the average f_i value in G_j, and $var(f_{ij})$ the variance of the f_i in G_j. The following formula, inspired by ANOVA (analysis of variance), gives an indication of how relevant f_i is for the function set as a whole:

$$A_i = \frac{Var_j[\bar{f}_{ij}]}{Mean_j[var(f_{ij})]} \tag{1}$$

where Var_j and $Mean_j$ denote the Variance and Mean operators taken over all values of j. A high A_i denotes a high relevance of feature f_i.

In the following, we will empirically determine whether the shortest-path distances to all, or a few particular, nodes are indeed informative with respect to a protein's functions by evaluating the performance of the method on a benchmark dataset.

5 Experiments

We performed two consecutive experiments. Firstly, we evaluated the potential of the proposed protein description for protein function prediction by assessing multiple learning systems and finding the learning system whose inductive bias best fits our dataset. This step was made to alleviate the risk of concluding that the description is insuitable, when the cause for bad results is in fact a poor choice of learner. Secondly, we compared the performance of this system with that of the Majority Rule, a transductive learner.[4]

We evaluate predictive performance using the following measures: area under the ROC curve (AUC) [10], precision, recall, and F1. We do not include predictive accuracy (percentage of predictions that are correct) because for several function prediction tasks, the class distribution is highly skewed (e.g., 1% of the protein has that function, 99% does not), and in such cases predictive accuracy (the percentage of predictions that is correct) does not carry much information. AUC and precision/recall are much more robust to skewed class distributions.

5.1 Datasets

We apply our method to four S.crevisiae PPI networks which are DIP-Core [11], VonMering [12], Krogan [13] and MIPS [14]. DIP-Core, VonMering, Krogan and MIPS have 4400, 22000, 14246 and 44514 interactions among 2388, 1401, 2708 and 7928 proteins respectively. We consider 18 high level functions for evaluating our function predictors.

5.2 Comparison of Learners

Given the input data and a particular function to predict, any standard machine learning tool can be used to build a model that predicts from a node's description whether the node has a particular function or not. We have experimented with several methods, as available in the Weka data mining toolbox [15], namely decision trees (J48), random forests, an instance based learner (IBk), Naive Bayes, radial basis function networks, Support Vector Machine (SMO), Classification via Regression (CVR) and Voting Feature Intervals (VFI). These methods were chosen to be representative for a broad range of machine learning methods. This

[4] Majority Rule was selected for its ease of implementation, and because it is still a regularly used reference method.

comparative evaluation was made on the DIP data set. The results are shown in Figure 1. RF performs best among all learners in 14 out of 18 cases. The 4 cases where it does not are all characterized by a very high class skew. The latter is not so surprising: Random Forests are ensembles of decision trees, and these are known to perform less well on highly skewed class distributions.

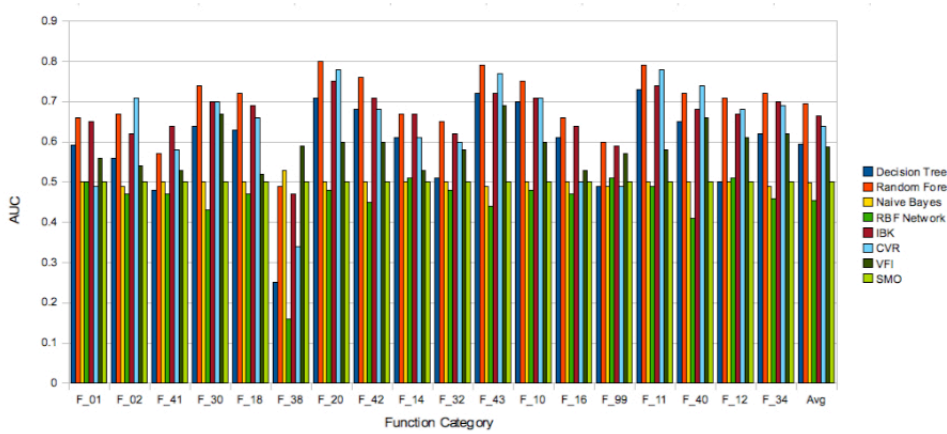


Fig. 1. Comparison of different machine learning methods on the DIP Dataset.

We have concluded from these results that the Random Forests method is our best candidate for learning from this type of data, and we have used this method in the remaining experiments.

5.3 Comparison with a transductive method

We next compare Random Forests and Majority Rule in predicting the proteins functions of four datasets DIP-Core, VonMering, Krogan and MIPS. Firstly, we select 700 nodes based on the Anova Measure. Then, we found the shortest path of each protein to those selected proteins. We used this information as the input for Weka and calculated the average Precision, Recall, F-Measure and AUC for each function class in a 10-fold cross validation. Figure 2 compares the average precision, over all classes, of Majority Rule (MR) and Random Forests (RF). Figure 3 similarly compares the recall of MR and RF, and Figure 4 the F1-measures. We see that, over the four datasets, RF has higher precision (11% higher in average) but smaller Recall (10% smaller in average). RF and MR perform almost similarly with respect to Fmeasure. The AUCs are compared in Figure 5 ; again, RF tends to have higher scores (+6%)

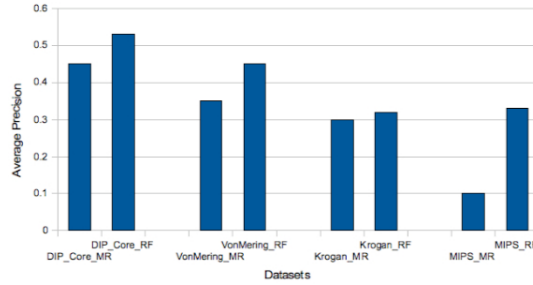


Fig. 2. Average precision of MR and RF in four datasets.

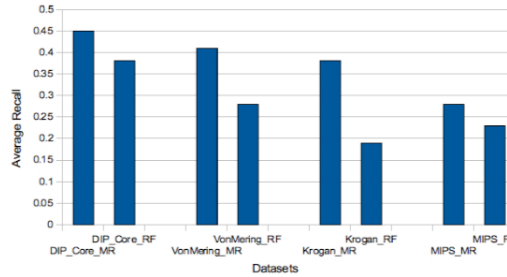


Fig. 3. Average recall of MR and RF in four datasets.

6 Conclusions

To summarize, we have firstly classified existing methods for the prediction of node properties in a network into transductive and inductive methods; this distinction provides insight in potential strengths and weaknesses of the methods, particularly in terms of the bias of the learning method. Inductive learning methods make different assumptions about the true labeling function than transductive methods which helped us in our choice of algorithm employed in this work. Secondly, we observed that existing inductive learning methods for predicting protein functions in PPI networks use local information, while the use of global information for such methods has as of yet remained unexplored. Accordingly, we have, thirdly, introduced a node description formalism that has not been used previously for protein function prediction and which is global. On four benchmark datasets, DIP-Core, VonMering, Krogan and MIPS, we have shown that this method outperforms the benchmark Majority Rule approach according to Precision and AUC and, hence, that it is informative with respect to the prediction of the function of a protein from the functions of its neighbors.

In the future, a more extensive comparison with other learners would be warranted. It would also be interesting to determine to what extent the information in our global protein description is complementary to that used in other (local inductive, or transductive) approaches. The reason is that when several predictors exploit different information when making their predictions, they can

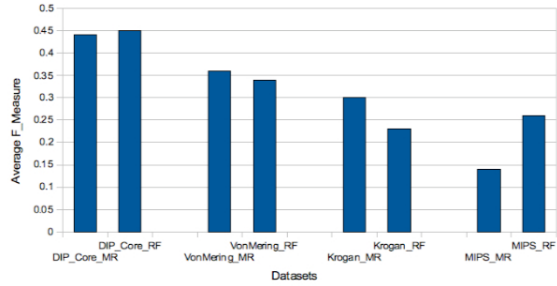


Fig. 4. Average f-Measure of MR and RF in four datasets.

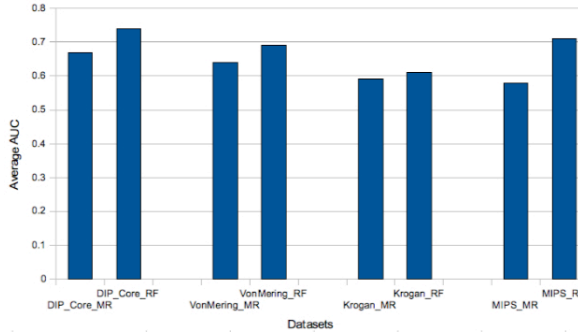


Fig. 5. Average AUC of MR and RF in four datasets.

typically be combined into a single composite predictor that performs better than each individual one. Finally, while we have focused here on models that predict a single class at a time, there exist a few methods that predict multiple classes simultaneously [16]; it would be useful to investigate to what extent these classifiers yield better predictions than the single-label prediction approach presented here.

References

1. Mitchell, T.: Machine Learning. McGraw-Hill Education (ISE Editions) (1997)
2. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nat Biotechnol **18** (2000) 1257–1261
3. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. Nat Biotechnol **21** (2003) 697–700
4. Sun, S., Zhao, Y., Jiao, Y., Yin, Y., Cai, L., Zhang, Y., Lu, H., Chen, R., Bu, D.: Faster and more accurate global protein function assignment from protein interaction networks using the mfgo algorithm. FEBS Lett **580** (2006) 1891–1896

5. Chua, H.N., Sung, W.K., Wong, L.: Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics **22** (2006) 1623–1630
6. King, A.D., Pržulj, N., Jurisica, I.: Protein complex prediction via cost-based clustering. Bioinformatics **20** (2004) 3013–3020
7. Brun, C., Herrmann, C., Guénoche, A.: Clustering proteins from interaction networks for the prediction of cellular functions. BMC Bioinformatics **5** (2004) 95
8. Milenkovic, T., Przulj, N.: Uncovering biological network function via graphlet degree signatures. (2008)
9. Rives, A.W., Galitski, T.: Modular organization of cellular networks. Proceedings of the National Academy of Sciences **100** (2003) 1128–1133
10. Provost, F.J., Fawcett, T.: Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: KDD. (1997) 43–48
11. Deane, C.M., Salwiński, Ł., Xenarios, I., Eisenberg, D.: Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics **1** (2002) 349–356
12. Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417** (2002) 399–403
13. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A., Greenblatt, J.F.: Global landscape of protein complexes in the yeast saccharomyces cerevisiae. Nature **440** (2006) 637–643
14. Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schüller, C.M., Stocker, S., Weil, B.: Mips: a database for genomes and protein sequences. Nucleic Acids Research **28** (2000) 37–40
15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (1999)
16. Struyf, J., Dzeroski, S., Blockeel, H., Clare, A.: Hierarchical multi-classification with predictive clustering trees in functional genomics. In: EPIA. (2005) 272–283

Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction

Matteo Re and Giorgio Valentini

DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano,
Via Comelico 39, 20135 Milano, Italia.
{re,valentini}@dsi.unimi.it

Abstract. Several works showed that biomolecular data integration is a key issue to improve the prediction of gene functions. Quite surprisingly only little attention has been devoted to data integration for gene function prediction through ensemble methods. In this work we show that relatively simple ensemble methods are competitive and in some cases are also able to outperform state-of-the-art data integration techniques for gene function prediction.

1 Introduction

The availability of an ever increasing amount of data sources due to recent advances in high throughput biotechnologies opens unprecedented opportunities for genome-wide gene function prediction. Indeed several works showed that biomolecular data integration play an essential role in the prediction of genes/gene products functions.

Gene function prediction in its general formulation is a complex classification problem characterized by the following items: a) each gene/gene product can be assigned to multiple terms/classes (a multiclass, multilabel classification problem); b) classes are structured according to a predefined hierarchy (a directed acyclic graph for the Gene Ontology [1] or a tree forest for FunCat [2]); c) classes are usually unbalanced (with positive examples usually less than negatives); d) known gene labels are in several cases be uncertain; e) multiple sources of data can be used to predict gene functions.

In this paper we focus on the last item, considering the problem of the prediction of a subset of FunCat classes in the model organism *S. cerevisiae*.

The main approaches proposed in the literature can be schematically subdivided in three categories: functional linkage networks, vector subspace integration and kernel fusion methods [3]. Modelling interactions between gene products using functional linkage networks is realized through graphs, where gene products are modeled as nodes and relationships between genes through edges [4]. In vector space integration (VSI) different vectorial data are concatenated [5], while kernel methods, by exploiting the closure property with respect to the

sum or other meaningful algebraic operators represent another valuable research direction for the integration of biomolecular data [6].

All these methods suffer of limitations and drawbacks, due to their limited scalability to multiple data sources (e.g. Kernel integration methods based on semidefinite programming [6]), to their limited modularity when new data sources sources are added (e.g. vector-space integration methods), or when data are not available as relational data (e.g. functional linkage networks).

Quite surprisingly, as observed by William Noble and Asa Ben-Hur [3], only little attention has been devoted to ensemble methods as a mean to integrate multiple biomolecular sources of data for gene function prediction. To our knowledge only few works very recently considered ensemble methods in this specific bioinformatics context: Naive-Bayes integration of the outputs of SVMs trained with multiple sources of data [7], and logistic regression for combining the output of several SVMs trained with different data and kernels in order to produce probabilistic outputs corresponding to specific GO terms [8].

The main aim of this work consists in showing that simple ensemble methods can obtain results comparable with state-of-the-art data integration methods, exploiting at the same time the modularity and scalability that characterize most of the ensemble algorithms. Indeed biomolecular data differing for their structural characteristics (e.g. sequences, vectors, graphs) can be easily integrated, because with ensemble methods the integration is performed at the decision level, combining the outputs produced by classifiers trained on different datasets. Moreover, as new types of biomolecular data, or updates of data contained in public databases, are made available to the research community, ensembles of learning machines are able to embed new data sources or to update existing ones by training only the base learners devoted to the newly added or updated data, without retraining the entire ensemble. In other words ensemble methods scale well with the number of the available data sources, and problems that characterize other data fusion approaches are thus avoided.

2 Methods

2.1 Ensemble methods

Data fusion can be realized by means of an ensemble system composed by learners trained on different "views" of the data and then combining the outputs of the component learners. Each type of data may capture different and complementary characteristics of the objects to be classified and the resulting ensemble may obtain better prediction capabilities through the diversity and the anti-correlation of the base learner responses.

We programmatically considered simple methods:

Weighted majority voting [10], using linear or logarithmic weights, tuned on the F-measure estimated from the training data, since gene functional classes are usually unbalanced.

Naive Bayes : a combination of classifiers assuming independence between them, that estimates the class-conditional support given the observed vector of categorized component classifiers outputs [11].

Decision Templates : a combination method based on the comparison of a "prototypical answer" of the ensemble for the examples belonging to a given class (the template) with the current answer of the ensemble to a specific example whose class needs to be predicted (the decision profile) [12].

The decision profile $\mathrm{DP}(\mathbf{x})$ for an instance $\mathbf{x}$ is a matrix composed by $d_{t,j} \in [0,1]$ elements representing the support (e.g. the probability) given by the t^{th} classifier to class ω_j. Decision templates DT_j are the averaged decision profiles obtained from $\mathbf{X}_j$, the set of training instances belonging to the class ω_j:

$$DT_j = \frac{1}{|\mathbf{X}_j|} \sum_{\mathbf{x} \in \mathbf{X}_j} DP(\mathbf{x}) \tag{1}$$

By computing the similarity $\mathcal{S}$ between $DP(\mathbf{x})$ and the decision template DT_j for each class ω_j, from a set of c classes, the final decision of the ensemble is taken by assigning a test instance $\mathbf{x}$ to a class with the largest similarity [12]:

$$D(\mathbf{x}) = \arg\max_j \mathcal{S}_j(\mathbf{x}) \tag{2}$$

It is easy to see that with dichotomic problems the decision templates are reduced to two-columns matrices, and the similarity ($\mathcal{S}_1$) for the positive class and the similarity ($\mathcal{S}_2$) for the negative class can be computed as 1 minus the normalized squared euclidean distance:

$$\mathcal{S}_1(\mathbf{x}) = 1 - \frac{1}{n} \sum_{t=1}^{n} [DT_1(t,1) - d_{t,1}(\mathbf{x})]^2 \tag{3}$$

$$\mathcal{S}_2(\mathbf{x}) = 1 - \frac{1}{n} \sum_{t=1}^{n} [DT_2(t,1) - d_{t,1}(\mathbf{x})]^2 \tag{4}$$

where DT_1 is the decision template for the positive and DT_2 for the negative class. The final decision of the ensemble is:

$$D(\mathbf{x}) = \arg\max_{\{1,2\}} (\mathcal{S}_1(\mathbf{x}), \mathcal{S}_2(\mathbf{x})) \tag{5}$$

2.2 Kernel fusion and vector space integration

Kernel fusion (KF) for data integration is based on the closure property of kernels with respect to the sum or other algebraic operators [6]. In our experiments we integrated the different data sets by simply summing their Gram matrices, and then we trained the SVMs directly with the resulting matrix. Vector space integration (VSI) consists in concatenating the vectors of the different data sets [5]. The resulting concatenated vectors are used to train a SVM. Note that training a linear SVM with concatenated vectors (VSI) is equivalent to kernel fusion with linear kernels. In our experiments we used gaussian kernels.

Table 1. Datasets

Code	Dataset	examples	features	description
D_{ppi1}	PPI - STRING	2338	2559	protein-protein interaction data from [13]
D_{ppi2}	PPI - BioGRID	4531	5367	protein-protein interaction data from the *BioGRID* database [14]
D_{pfam1}	Protein domain log-E	3529	5724	Pfam protein domains with log E-values computed by the *HMMER* software toolkit
D_{pfam2}	Protein domain binary	3529	4950	protein domains obtained from *Pfam* database [15]
D_{expr}	Gene expression	4532	250	merged data of Spellman and Gasch experiments
D_{seq}	Pairwise similarity	3527	6349	Smith and Waterman log-E values between all pairs of yeast sequences

3 Experimental results

Even if the growing rate of the amount of biomolecular data available for many species was constantly increasing in the last years, the model organisms with a consistent amount of literature inherent to data fusion based gene function prediction are actually reduced to S.*cerevisiae* and M.*musculus*. Despite the availability of a well established public benchmark dataset, such as the one provided during the MouseFunc contest [18], a recent comparison between many model organisms showed that the fraction of genes annotated with experimental evidence is about 30% larger in S.*cerevisiae* than in M.*musculus* (85.4% and 57.8% respectively for the yeast and mouse model organisms) [19]. We thus decided to use yeast data for our experiments. In order to maximize the effective use of the larger experimental coverage of gene functional annotations available for the yeast, we also adopted as a reference functional ontology, the MIPS Functional Catalogue (FunCAT), which is composed by annotations mainly based on experimental evidences [2], allowing us to minimize the impact of non experimental functional annotations.

We predicted the top-level 15 functional classes of the FunCat taxonomy of the model organism *S. cerevisiae*, using 6 different sources of data (Tab. 1). Each dataset was split into a training set and a test set (composed, respectively, by the 70% and 30% of the available samples), considering yeast genes common to all data sets (about 1900) and with at least 1 FunCat annotation. A 3-fold stratified cross-validation has been performed on the training data for model selection, using gaussian SVMs with probabilistic output [9] as base learners for ensemble methods, and for VSI and KF data integration. We compared the performances of single gaussian SVMs trained on each data set with those obtained with vector-space-integration (VSI) techniques, kernel fusion through the sum of gaussian kernels, and with the ensembles described in Sect. 2.1.

Table 2 shows the average F-measure, recall, precision and AUC across the 15 selected FunCat classes, obtained through the evaluation of the test sets (each constituted by 570 genes). The four first columns refer respectively to the weighted linear, logarithmic linear, decision template and naive Bayes ensembles;

VSI and KF stands respectively for vector space integration and kernel fusion, D_{avg} represents the average results of the single SVMs across the six datasets, and D_{ppi2} represents the single SVM that achieved the best performance, i.e. the one trained using protein-protein interactions data collected from BioGrid. Tab. 3 shows the same results obtained by each single SVM trained on a specific biomolecular data set.

Looking at the values presented in Tab. 2, on the average, data integration through simple ensemble methods provide better results than single SVMs, VSI and Kernel fusion, independently of the applied combination rule. In particular, Decision Templates achieved the best average F-measure, and ensemble methods as a whole the best AUC. Among the ensemble of classifiers, with respect to the AUC, the worst performing method is the Naive Bayes combiner albeit its performances are still, on the average, higher than the ones reported for VSI, Kernel fusion and the single classifiers. Precision of the ensemble methods is relatively high: this is of paramount importance to drive the biological validation of "in silico" predicted functional classes: considering the high costs of biological experiments, we need to obtain a high precision (and possibly recall) to be sure that positive predictions are actually true with the largest confidence.

To understand whether the differences between AUC scores in the 15 dichotomic tasks are significant, we applied a non parametric test based on the Mann-Whitney statistic [16], using a recently proposed software implementation [17]. Tab. 4 shows that at 0.01 significance level in most cases there is no significant difference between AUC scores of the weighted linear and logarithmic ensembles (E_{lin} and E_{log}) and the Decision Template (E_{dt}) combiner. A different behavior is observed for the Naive Bayes combiner: its performances are comparable to the ones obtained by the other ensemble methods only in 2 over 15 classification tasks and worse in the remaining 13.

Most interestingly, ensemble methods significantly outperform the other data integration methods. For instance, wins-ties-losses of E_{lin} vs VSI are $13-2-0$, and $9-6-0$ vs KF; Naive-Bayes, the worst performing ensemble method, achieves $9-6-0$ wins-ties-losses with VSI and $5-10-0$ with KF. It is worth noting that, among the tested ensemble methods, E_{lin}, E_{log} and E_{dt} undergo no losses when compared with single SVMs (Tab. 4, bottom): we can safely choose any ensemble method (but not the Naive Bayes combiner) to obtain equal or

Table 2. Ensemble methods, kernel fusion and vector space integration: average F-score, recall, precision and AUC (Area Under the Curve) across the data sets.

Metric	E_{lin}	E_{log}	E_{dt}	E_{NB}	VSI	KF	D_{avg}	D_{ppi2}
F	0.4347	0.4111	0.5302	0.5174	0.3213	0.3782	0.3544	0.4818
rec	0.3304	0.2974	0.4446	0.6467	0.2260	0.3039	0.2859	0.3970
prec	0.8179	0.8443	0.7034	0.5328	0.6530	0.6293	0.5823	0.6157
AUC	0.8642	0.8653	0.8613	0.7933	0.7238	0.7775	0.7265	0.8170

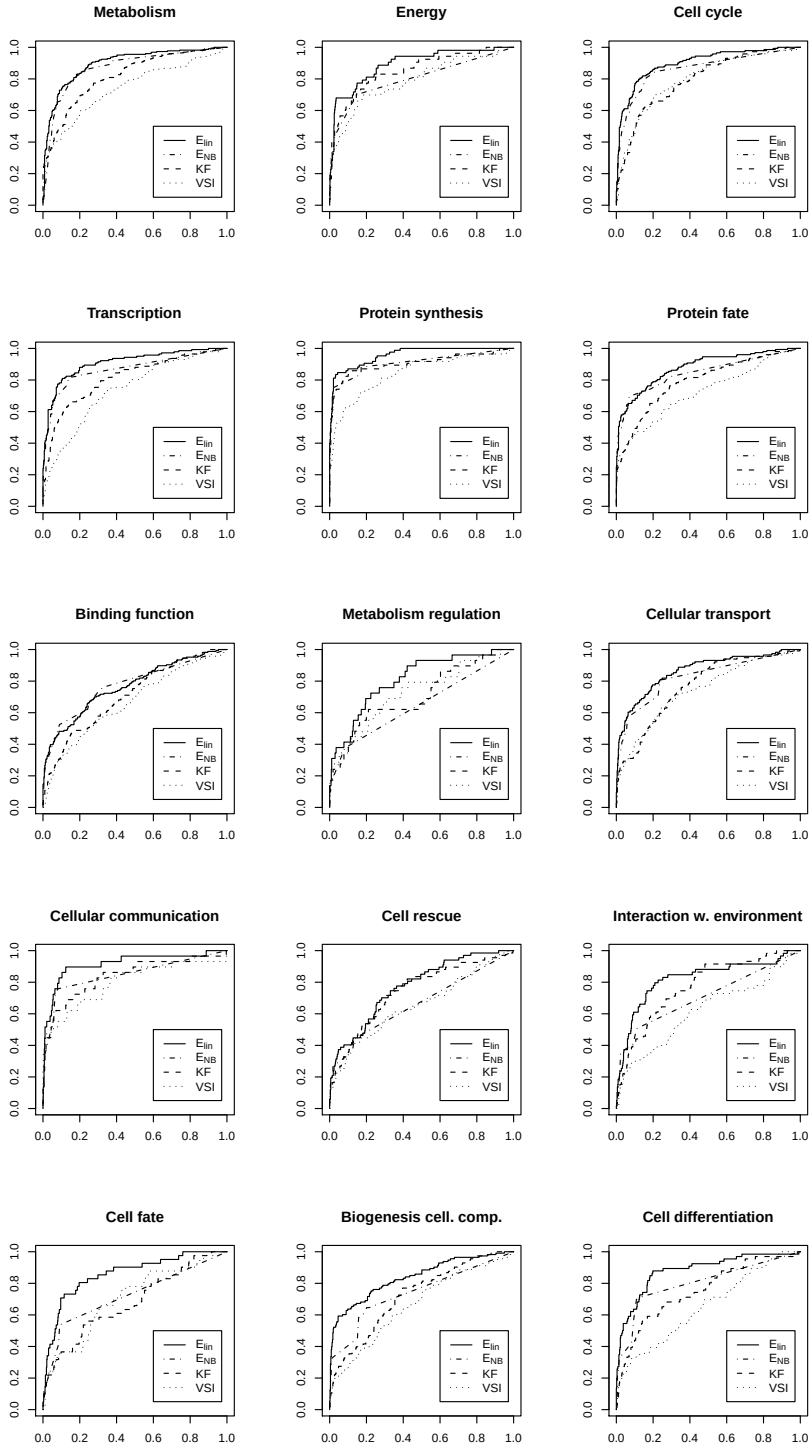


Fig. 1. Comparison of ROC curves between different data integration methods. E_{lin}: ensemble weighted majority voting; E_{NB}: Naive-Bayes ensemble integration; KF: kernel fusion; VSI: vector space integration.

Table 3. Single SVMs: average F-score, recall, precision and AUC. Each SVM is identified by the same name of the data set used for its training (Tab. 1).

Metric	D_{ppi1}	D_{ppi2}	D_{pfam1}	D_{pfam2}	D_{expr}	D_{seq}
F	0.3655	0.4818	0.2363	0.3391	0.2098	0.4493
rec	0.2716	0.3970	0.1457	0.2417	0.1571	0.5019
prec	0.6157	0.6785	0.7154	0.6752	0.3922	0.4162
AUC	0.7501	0.8170	0.6952	0.6995	0.6507	0.7469

Table 4. Results of the non-parametric test based on Mann-Whitney statistics to compare AUCs between ensembles, VSI, Kernel fusion and single SVMs. Each entry represents wins-ties-losses between the corresponding row and column at 0.01 significance level. Top: Comparison between ensemble methods, VSI and kernel fusion; Bottom: Comparison between data integration methods and single SVMs.

	VSI	E_{log}	E_{lin}	E_{dt}	E_{NB}
E_{log}	13-2-0	-	-	-	-
E_{lin}	13-2-0	0-14-1	-	-	-
E_{dt}	13-2-0	1-13-1	1-11-3	-	-
E_{NB}	9-6-0	0-2-13	0-2-13	0-2-13	-
KF	3-12-0	0-6-9	0-6-9	0-6-9	0-10-5

	D_{ppi1}	D_{ppi2}	D_{pfam1}	D_{pfam2}	D_{expr}	D_{seq}
E_{lin}	11-4-0	4-11-0	15-0-0	14-1-0	15-0-0	13-2-0
E_{log}	11-4-0	4-11-0	15-0-0	14-1-0	15-0-0	13-2-0
E_{dt}	11-4-0	4-11-0	15-0-0	14-1-0	15-0-0	13-2-0
E_{NB}	5-10-0	2-11-2	9-6-0	8-7-0	12-3-0	7-8-0
VSI	1-11-3	0-8-7	2-11-2	1-14-0	4-11-0	0-12-3
KF	1-14-0	0-9-6	5-10-0	5-10-0	11-4-0	3-12-0

better results than any of the single SVMs. On the contrary in many cases VSI, E_{NB} and the kernel fusion methods obtained worse results than single SVMs, although performances achieved by the Naive Bayes combiner and the kernel fusion methods are, in general, better than those obtained by VSI. Nevertheless, we can observe that a single SVM trained with Ppi-2 data achieves good results (11 ties with ensembles and an average AUC $\simeq 0.81$ w.r.t. 0.86 of the ensembles, Tab. 2 and 4), showing that large protein-protein interactions data sets alone provide information sufficient to correctly predict several FunCat classes.

Fig. 1 compares the ROC curves of the different data integration methods used in our experiments. ROC curves of weighted majority voting (E_{lin}) are consistently above the corresponding ROC curves of kernel fusion and vector space integration for all the considered FunCat classes. ROC curves of Naive Bayes combiner are below those of kernel fusion only for four classes: "Energy", "Metabolism", "Regulation", "Cell rescue" and "Interaction with the environment".

4 Conclusions

The main objective of this contribution is to demonstrate that simple ensemble methods are competitive with state-of-the-art methods for gene function prediction based on heterogeneous biomolecular data integration.

It is well-known that gene function prediction methods need to take into account the hierarchical relationships between classes to improve their predictions [7, 8, 20]. Nevertheless, in this investigation we focused on data integration, in order to study the improvement due to the usage of multiple sources of data, without exploiting any knowledge about the hierarchical relationships between classes. In this way we can separate the contribution due to data fusion techniques from the improvement due to hierarchical methods.

Considering the increasing growing rate of available biomolecular data, the modularity and scalability that characterize ensemble methods can favour an easy update of existing sources of data and an easy integration of new ones. Our preliminary experiments show that relatively simple ensemble methods are competitive with kernel fusion and vector space integration, two of the most largely applied machine learning data integration techniques for gene function prediction. This could seem quite surprisingly, but considering the uncertainty that characterize both annotations and measurements of data values, we can expect that relatively simple methods are able to nicely work in a similar context. Moreover it is worth noting that each type of data can only capture a particular characteristic of a protein, and for different functional classes the same type of data can be highly informative or completely unuseful to discriminate positive and negative examples. For these reasons the inherent modularity and adaptivity of ensemble systems can explain their effectiveness for the integration of multiple biomolecular data sources. In particular we think that ensemble methods devoted to biomolecular data integration can be a valuable research line to improve the accuracy of gene function prediction problems.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments and suggestions. The authors gratefully acknowledge partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views.

References

[1] The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. Nature Genet. **25** (2000) 25–29

[2] Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., Mewes, H.: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research **32** (2004) 5539–5545

[3] Noble, W., Ben-Hur, A.: Integating information for protein function prediction. In Lengauer, T., ed.: Bioinformatics - From Genomes to Therapies. Volume 3. Wiley-VCH (2007) 1297–1314
[4] Karaoz, U., et al.: Whole-genome annotation by using evidence integration in functional-linkage networks. Proc. Natl Acad. Sci. USA **101** (2004) 2888–2893
[5] desJardins, M., Karp, P., Krummenacker, M., Lee, T., Ouzounis, C.: Prediction of enzyme classification from protein sequence without the use of sequence similarity. In: Proc. of the 5th ISMB, AAAI Press (1997) 92–99
[6] Lanckriet, G., De Bie, T., Cristianini, N., Jordan, M., Noble, W.: A statistical framework for genomic data fusion. Bioinformatics **20** (2004) 2626–2635
[7] Guan, Y., Myers, C., Hess, D., Barutcuoglu, Z., Caudy, A., Troyanskaya, O.: Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology **9** (2008)
[8] Obozinski, G., Lanckriet, G., Grant, C., M., J., Noble, W.: Consistent probabilistic output for protein function prediction. Genome Biology **9** (2008)
[9] Lin, H., Lin, C., Weng, R.: A note on Platt's probabilistic outputs for support vector machines. Machine Learning **68** (2007) 267–276
[10] Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence **20** (1998) 226–239
[11] Titterington, D., Murray, G., Spiegelhalter, D., Skene, A., Habbema, J., Gelpke, G.: Comparison of discriminant techniques applied to a complex data set of head injured patients. Journal of the Royal Statistical Society **144** (1981)
[12] Kuncheva, L., Bezdek, J., Duin, R.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition **34** (2001) 299–314
[13] vonMering, C., et al.: STRING: a database of predicted functional associations between proteins. Nucleic Acids Research **31** (2003) 258–261
[14] Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. **34** (2006) D535–D539
[15] Finn, R., Tate, J., Mistry, J., Coggill, P., Sammut, J., Hotz, H., Ceric, G., Forslund, K., Eddy, S., Sonnhammer, E., Bateman, A.: The Pfam protein families database. Nucleic Acids Research **36** (2008) D281–D288
[16] Delong, E., Delong, D., Clarke-Pearson, D.: Comparing the areas under two or more or more correlated Receiver Operating Characteristics Curves: a non parametric approach. Biometrics **44** (1988) 837–845
[17] Vergara, I., Norambuena, T., Ferrada, E., Slater, A., Melo, F.: StAR: a simple tool for the statistical comparison of ROC curves. BMC Bioinformatics **9** (2008)
[18] Pena Castillo, L., et al.: A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biology **9:S2** (2008)
[19] Rhee, S.Y., et al.: Use and misuse of the gene ontology annotations. Nature Rev. Genetics **9** (2008) 509–515
[20] Valentini, G. and Re, M.: Weighted True Path Rule: a multilabel hierarchical algorithm for gene function prediction. In: MLD-ECML 2009, 1st International Workshop on learning from Multi-Label Data, Bled, Slovenia (2009) 133–146

Integrated network construction using event based text mining

Yvan Saeys, Sofie Van Landeghem, and Yves Van de Peer

Department of Plant Systems Biology, VIB
Department of Molecular Genetics, University of Ghent
9052, Gent, Belgium
{yvan.saeys,sofie.vanlandeghem,yves.vandepeer}@psb.ugent.be

Abstract. The scientific literature is a rich and challenging data source for research in systems biology, providing numerous interactions between biological entities. Text mining techniques have been increasingly useful to extract such information from the literature in an automatic way, but up to now the main focus of text mining in the systems biology field has been restricted mostly to the discovery of protein-protein interactions. Here, we take this approach one step further, and use machine learning techniques combined with text mining to extract a much wider variety of interactions between biological entities. Each particular interaction type gives rise to a separate network, represented as a graph, all of which can be subsequently combined to yield a so-called integrated network representation. This provides a much broader view on the biological system as a whole, which can then be used in further investigations to analyse specific properties of the network.

1 Introduction

A wealth of biological information is currently recorded in scientific publications, which are easily accessible through online literature services like PubMed[1]. However, such resources are expanding exponentially and in order to keep up with the recent literature and retrieve relevant biological information, automated systems have become a time saving necessity.

Text mining methods are data mining techniques that focus on extracting relevant knowledge from these largely unstructured texts. Their use in systems biology started with simple, co-occurrence based methods that suggested relations between entities when they appeared in the same sentence [Ding et al, 2002], typically exhibiting high recall, but low precision [Hoffmann and Valencia, 2004]. As high precision frameworks are often preferred in systems biology, especially when integrating different data sources, more elaborated techniques, either based on hand-crafted rules [Fundel et al, 2007] or machine learning methods have been introduced. We will focus here on the latter techniques as they scale better to

[1] http://pubmed.gov

large datasets, and can be easily retrained when more data becomes available.

Up to now, the main focus of text mining techniques that rely on machine learning approaches has been the automatic extraction of protein-protein interactions, or the association of genes to certain diseases. A number of evaluation corpora have been built to assess the performance of techniques on the first of these tasks [Pyysalo et al, 2008,Van Landeghem et al, 2008a]. Recently, the BioNLP'09 shared task was initiated as a community-wide effort to leverage the scope of text mining techniques to extract more complex events from text, in order to capture a wider variety of interactions and thus gain more knowledge from information encoded in the literature [Kim et al., 2009].

The main task in this challenge was to identify as good as possible 9 different types of bio-molecular events. For each event, the organizers provided a set of annotated PubMed abstracts, which could be used by the participants to train their models. Afterwards, a separate validation set was provided, allowing participants to evaluate their predictions, and finally an independent test set was provided to which all participants were evaluated.

In this work, we describe a machine learning approach that uses graph-based features from sentence representations to detect these different types of interactions, and subsequently uses them to construct an integrated network that contains all high-confidence predictions. The remainder of the manuscript is structured as follows. First we elaborate on the methodology we used to convert these problems into a machine learning setting, outlining the general preprocessing of the documents, the applied machine learning techniques, and the final postprocessing to ensure a high-precision approach. Next, we present the results of this analysis: the evaluation of the whole framework on the BioNLP'09 evaluation and test set, and the construction of an integrated network using these predictions. We conclude by highlighting future perspectives and challenges that remain in this domain.

2 Methods

The core part of the BioNLP'09 challenge concerned the automatic detection and characterization of bio-molecular events from text. There are 9 distinct event types, six of which influence proteins directly, further referred to as 'Protein events', and three which describe 'Regulation events'. Five of the protein events are unary: Localization, Gene expression, Transcription, Protein catabolism and Phosphorylation. The sixth protein event, Binding, can be either related to one protein (e.g. protein-DNA binding), two proteins (e.g. protein-protein interaction) or more (e.g. a complex). The three types of Regulation events are the following: Regulation (unspecified), Positive regulation and Negative regulation. Each of them can be unary or binary. In the latter case, an extra argument specifying the cause of the regulation is added. Each argument of a Regulation event can be either a protein or any other event. This offers opportunities to

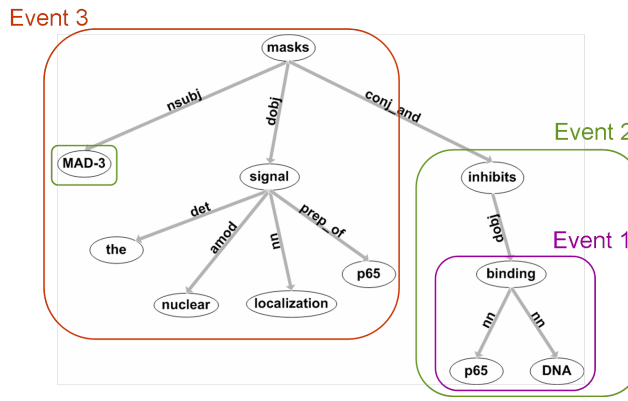


Fig. 1. Example of a dependency graph for the sentence 'MAD-3 masks the nuclear localization signal of p65 and inhibits p65 DNA binding'. The three events represented in this sentence are indicated in the respective subgraphs.

detect different levels of interactions, and thus detect Regulation events in an iterative way.

The detection of Protein and Regulation events can now be stated as a set of binary classification problems, one for each event. A given potential occurrence of an event should then be scored by a classification model, which would either accept or reject the current example as being an instance of the particular event type. We will now go into more detail on how to transform the unstructured text data into a well defined classification task.

2.1 Data preprocessing

A challenging problem in text mining is to find an appropriate representation of the text, allowing machine learning techniques to make use of features that represent the key information to solve the task at hand. A few steps should be performed in order to transform the data into such a useful format.

In a first step, informative sentences containing biological entities are selected (information retrieval), and those key entities are identified and tagged in the sentence (named entity recognition). Subsequently, a deep syntactic parsing of each sentence was performed using the Stanford parser [de Marneffe et al, 2006], resulting in part-of-speech tags and dependency graphs. A dependency graph models the syntactic structure of a sentence, and is often used in many machine learning approaches as a structured data type to be used as input for the classification model [Zelenko et al, 2008,Kim et al, 2008].

Figure 1 shows an example of a dependency graph for the sentence 'MAD-3 masks the nuclear localization signal of p65 and inhibits p65 DNA binding'. This sentence contains three events to be detected by the system: 1) a Binding

Event type	# Features	**# neg. inst.**	**# pos. inst.**	**% pos. inst.**
Localization	18 121	3415	249	7
Single binding	21 332	3548	522	13
Multiple binding	11 228	2180	185	8
Gene expression	31 332	5356	1542	22
Transcription	30 306	6930	489	7
Protein catabolism	1 883	175	96	35
Phosphorylation	2 185	163	153	48
Unspecified regulation (Unary)	27 915	6076	408	6
Positive regulation (Unary)	48 944	13834	1367	9
Negative regulation (Unary)	16 673	3233	489	13
Unspecified regulation (Binary)	4 239	778	81	9
Positive regulation (Binary)	19 468	5405	249	4
Negative regulation (Binary)	4 166	819	29	3

Table 1. Statistics of the training data set.

event (p65 DNA binding), 2) a Negative Regulation event (MAD-3 masks the nuclear localization signal of p65) and 3) a higher level Negative regulation event (MAD-3 inhibits p65 DNA binding), where one of the arguments is a protein (MAD-3) and the other is an event in itself (p65 DNA binding).

To couple the words occurring in a sentence to a particular event, dictionaries of trigger words associated to each event were used (e.g. 'interaction' for Binding and 'secretion' for Localization). From the training data, we automatically compiled such dictionaries of triggers for each event type, applying the Porter stemming algorithm [Porter, 1980] to each trigger. This resulted in some entries in the dictionaries which were of limited use, such as 'through' for Binding, or 'are' for Localization. Such words are too general or too vague, and will lead to many negative and irrelevant instances. For this reason, we manually cleaned the dictionaries, only keeping specific triggers for each event type.

2.2 Model setup

To extract useful features from the dependency graph, we used a rich feature representation based on our earlier work on predicting protein-protein interactions [Van Landeghem, 2008b]. The feature sets are a combination of information derived from the dependency tree (such as properties of the subgraph covering the event and lexical information of the trigger words) and information concerning the occurrence of words in the subgraph. The following features were extracted:

- A bag-of-words (BOW) approach which looks at all the words that appear at a vertex of the subgraph. This automatically excludes uninformative words such as prepositions. Here we used stemmed trigrams (succesions of three words) as BOW features.

- Lexical and syntactic information of triggers (stemmed versions of each word, as well as the associated part-of-speech tag generated by the parser).
- Size of the subgraph.
- Length of the sub-sentence.
- Extra features for Regulation events, storing whether the arguments are proteins or events, and specifying the exact event type.
- Vertex walks which consist of two vertices and their connecting edge. For these patterns, again lexical as well as syntactic information is kept. When using lexical information, protein names and triggers were blinded in order to extract more general patterns (e.g. 'trigger nsubj protx' which expresses that the given protein is the subject of a trigger).

The resulting datasets are inherently high-dimensional and very sparse. Table 1 shows the statistics of the training set for all event types. To deal well with these sparse, high-dimensional and class imbalanced datasets, SVMs are a natural choice for the classification model [Boser et al, 1992]. We used the LibSVM implementation of WEKA for our experiments, using the radial basis function (RBF) kernel as a default. As we were confronted with a separate validation and test set, only an internal 5-fold crossvalidation loop on the training data was used to optimize the C-parameter of the SVM, and the classification performance on the validation and test sets were used to assess model performance.

Finally, a number of custom-made post-processing modules were applied to the resulting predictions, aiming to further reduce false positives and hence improve the precision of our method. These include removing the weakest predictions if multiple events were predicted for the same trigger word, as well as reducing the number of predictions based on overlapping trigger words.

2.3 Integrated network construction

We take a graph based approach to combine the predictions of the different Protein and Regulation events. Consider a set of interaction events $\{I_1, I_2, \cdots, I_N\}$ to integrate into a network. We can then associate to each of the events I_i a graph G_i, obtained using the predictions of the SVM model for event I_i. Note that there exists a heterogeneity in the graphs, as there might be multiple edges between two nodes in a graph (due to more than one prediction for a certain edge), and that some of the edges may be directed (e.g. A regulates B) while others may be undirected (e.g binding of C and D). Furthermore, all edges are weighted by the confidence of the associated prediction (see further).

A convenient representation for each graph G_i is its associated matrix $G_i(jk)$ where each entry in the matrix is a *set* of weigthed connections between node j and node k. If there is no edge between node j and node k, then $G_i(jk) = \emptyset$. For undirected edges, the associated weight w_{jk} is represented both in $G_i(jk)$ and $G_i(kj)$, while for directed edges the weight is only added to the set representing the correct direction, this representation thus being a generalized form to combine both directed and undirected information.

Event type	Validation set			Test set		
	Recall	**Precision**	**F-score**	**Recall**	**Precision**	**F-score**
Localization	77.36	91.11	83.67	43.68	78.35	56.09
Binding	45.16	37.21	40.80	38.04	38.60	38.32
Gene expression	70.79	79.94	75.08	59.42	81.56	68.75
Transcription	60.98	75.76	67.57	39.42	60.67	47.79
Protein catabolism	80.95	89.47	85.00	64.29	60.00	62.07
Phosphorylation	68.09	88.89	77.11	56.30	89.41	69.09
Regulation	23.67	41.67	30.19	10.65	22.79	14.52
Positive regulation	21.56	38.00	27.51	17.19	32.19	22.41
Negative regulation	30.10	41.26	34.81	22.96	35.22	27.80

Table 2. Performance evaluation of all events for the validation and test datasets.

The weights on the edges are obtained by the classification model. For the SVM models, the distance to the hyperplane of each prediction is scaled between 0 and 1 such that the prediction threshold above which to decide on a positive prediction (this threshold varies per event) corresponds to a weight of 0.5.

It has to be noted that for some unary events, we may only know the effect, but not the causal node. In these cases, we introduce an artifical causal node for the effect node, which may be filled in later when more text is analysed. An integrated network can then be constructed by aggregating all matrices $G_i(jk)$ into a three-dimensional tensor $T(jkl)$ with dimensions $M \times M \times N$, where M is the cardinality of the union of all nodes in $G_i, i = 1 \cdots N$ and N is the number of events to integrate. The tensor entry $T(jkl)$ represents a connection from node j to node k for event type l. For visualisation purposes, we only keep all positive predictions, and discard all edges for which $T(jkl) < 0.5$.

3 Results

3.1 Predictive performance

To evaluate predictive performance, participants of the BioNLP'09 challenges could make use of a validation set to eventually finetune some parameters of their systems. However, performance could only be measured indirectly by submitting the predictions through a web interface, which then returned the evaluation measures (recall, precision and F-score). This only allowed for a rough, manual finetuning of some of the systems parameters, as an automatic exploration of parameter settings using this web interface was not possible. In our case, we only finetuned for each event the prediction threshold above which to consider a prediction to be positive.

Similarly, the final results on the test set were also assessed in a blind way: participants could only upload their predictions for this set one time, and after

Team	Protein Events	Binding	Regulation	All
UTurku	70.21	44.41	40.11	51.95
JULIELab	68.38	41.20	34.60	46.66
ConcordU	61.76	27.20	35.43	44.62
UT+DBCLS	63.12	31.19	32.30	44.35
VIBGhent	**64.59**	**38.32**	**22.41**	**40.54**
UTokyo	55.96	41.10	20.09	36.88
UNSW	55.39	28.92	20.90	34.92
UZurich	53.66	33.75	19.89	34.78
ASU+HU+BU	56.82	27.49	09.01	32.09
Cam	51.79	18.14	15.79	30.80

Table 3. Performance comparison for the top ten performing teams. Numbers shown denote the F-measure for the three types of events (columns Protein, Binding, and Regulation), as well as the overall performance (column All).

the submission deadline all evaluations were returned to the participants. Table 2 shows the evaluation measures for our system on both the validation (using optimized thresholds) and test set.

As can be expected, performance on the test set is lower than on the validation set, the decrease in F-measure ranging from only about 2% for Binding events, to 27% in the case of Localization events. In general, we achieve a high precision for Protein events: almost all results achieve a precision of 60% or more. Another trend is the fact that predicting Protein events achieves much higher performance than the prediction of Regulation events, a phenomenon that was observed by all participants in the challenge. This can be explained by the fact that the prediction of Regulation events largely depends on predicted Protein events (e.g. for higher level regulation events), thus causing false positives of predicted Protein events to cause even more false positive higher level regulation events.

To put these results into the context of the BioNLP'09 challenge, Table 3 compares the results of the ten best performing teams, out of 24 participating teams. Our team (VIBGhent) was ranked third for detecting Protein Events, fourth for detecting Binding Events, and fifth for detecting Regulation Events, resulting in an overall fifth ranking.

3.2 Constructing integrated networks

We created the tensor $T(jkl)$ for a set of six events $\{I_1, I_2, I_3, I_4, I_5, I_6\} = \{\text{Positive regulation}, \text{Negative regulation}, \text{Unspecified regulation}, \text{Binding}, \text{Transcription}, \text{Phosphorylation}\}$. Figure 2 shows a visualization of a subgraph of the integrated network, where the edge thickness corresponds to the prediction confidence of the interaction, and colors display different types of interactions

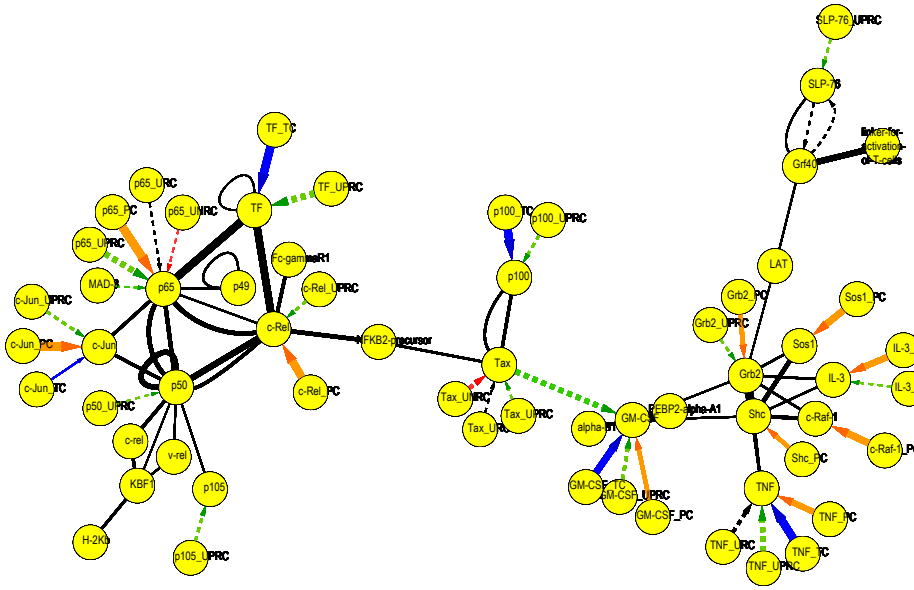


Fig. 2. Visualization of a subgraph of the integrated network, constructed on the combined results of the validation and the test set.

(black for Binding and unspecified Regulation events, orange for Phosphorylation, blue for Transcription and green/red for Positive/Negative Regulation events). Furthermore, Regulation events are displayed by dashed lines, and Protein events by full lines.

In a subsequent stage, the tensor $T(jkl)$ can be used to infer new biological knowledge, such as indirect interactions and pathways. An example of an indirect interaction, derived from the network depicted in Figure 2 is the positive regulation of GM-CSF by Tax, which is in turn negatively regulation by Tax_UNRC, which suggests an indirect regulation of GM-CSF by Tax_UNRC.

4 Conclusions and future work

In this work we presented a text mining approach that extracts various types of interactions from scientific literature. This information was used in a second stage to construct integrated networks, using the strength of the predictions as confidence weights for the connections in the network. As the application of text mining techniques for such problems is still in its childhood, improving the predictive performance of these techniques will remain a key challenge, as well as recognizing more adequately the specific type of interaction (e.g. protein-protein, protein-DNA, RNA-protein). Furthermore, we already performed some preliminary work on detecting speculation and negation of biological events, which will be useful to detect modes of (un)certainty about certain facts stated.

From a data integration point of view, we aim to combine the results obtained by text mining with interactions identified by other data sources (either experimentally verified or predicted) in order to increase the robustness of the networks.

5 Acknowledgements

Yvan Saeys and Sofie Van Landeghem would like to thank the Research Foundation Flanders (FWO-Vlaanderen) for funding their research. The authors would like to thank the anonymous reviewers for their suggestions to improve the mansucript.

References

[Boser et al, 1992] B. Boser, I. Guyon and V.N. Vapnik 1992. A training algorithm for optimal margin classifiers. *Proceedings of COLT 1992*, 144-152

[de Marneffe et al, 2006] MC. de Marneffe, B. MacCartney and C. D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*

[Ding et al, 2002] J. Ding, D. Berleant, D. Nettleton and E. Wurtele. 2002. Mining MEDLINE: abstracts, sentences, or phrases? *Proceedings of PSB'02*, :326-337

[Fundel et al, 2007] K. Fundel, R. Küffner and R. Zimmer. 2007. RelEx—Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365-371

[Hoffmann and Valencia, 2004] R. Hoffmand and A. Valencia. 2004. A Gene Network for Navigating the Literature. *Nature Genetics*, 36(7):664

[Kim et al, 2008] S. Kim, J. Yoon and J. Yang. 2008. Kernel approaches for genic interaction extraction. *Bioinformatics*, 24(1):118-126

[Kim et al., 2009] J.-D. Kim, T. Ohta, S. Pyssalo, Y. Kano and J. Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction, *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, to appear

[Porter, 1980] M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3), 130-137

[Pyysalo et al, 2008] S. Pyysalo, A. Airola, J. Heimonen, J. Bjrne, F. Ginter and T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl. 3):S6

[Van Landeghem et al, 2008a] S. Van Landeghem, Y. Saeys, B. De Baets and Y. Van de Peer. 2008. Benchmarking machine learning techniques for the extraction of protein-protein interactions from text. *Proceedings of the 18th Belgian Dutch Machine Learning Conference (Benelearn'08)*; 79-80.

[Van Landeghem, 2008b] S. Van Landeghem, Y. Saeys, B. De Baets and Y. Van de Peer. 2008. Extracting protein-protein interactions from text using rich feature vectors and feature selection. *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM)*, 77-84

[Van Landeghem, 2009] S. Van Landeghem, Y. Saeys, B. De Baets and Y. Van de Peer. 2009. Analyzing text in search of bio-molecular events: a high-precision machine learning framework. *Proceedings of BioNLP 2009*, 128-136

[Zelenko et al, 2008] D. Zelenko, C. Aone and A. Richardella. 2003. Kernel Methods for Relation Extraction. *JMLR*, 3(Feb):1083-1106

Evaluation Method for Feature Rankings and their Aggregations for Biomarker Discovery

Ivica Slavkov, Bernard Ženko, and Sašo Džeroski

Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
{Ivica.Slavkov,Bernard.Zenko,Saso.Dzeroski}@ijs.si
http://kt.ijs.si

Abstract. In this paper we investigate the problem of evaluating ranked lists of biomarkers, which are typically an output of the analysis of high-throughput data. This can be a list of probes from microarray experiments, which are ordered by the strength of their correlation to a disease. Usually, the ordering of the biomarkers in the ranked lists varies a lot if they are a result of different studies or methods. Our work consists of two parts. First, we propose a method for evaluating the "correctness" of the ranked lists. Second, we conduct a preliminary study of different aggregation approaches of the feature rankings, like aggregating rankings produced from different ranking algorithms and different datasets. We perform experiments on multiple public Neuroblastoma microarray studies. Our results show that there is a generally beneficial effect of aggregating feature rankings as compared to the ones produced by a single study or single method.

Key words: feature ranking evaluation, biomarker discovery, ranking aggregation

1 Introduction

In medicine, the progress or presence of some disease is determined by measuring certain biological parameters. These parameters are commonly called biomarkers and can range from blood pressure to the expression of a certain gene. Here, we focus on biomarkers derived from different types of high-throughput data.

We consider the process of biomarker discovery as the process of determining markers which have the strongest correlation to the presence or status of a certain disease. For example, given a microarray experiment, the output would be a list of probes ranked according to their differential expression. The main challenge in biomarker discovery from high dimensional data arises from having a small number of available biological samples, as well as from the inherent high variability of the data.

In machine learning terminology, biomarker discovery translates into the task of feature ranking and feature selection. Although these two tasks are related, they produce different result. On one hand, feature ranking provides an assessment of the "importance" of individual features to a target concept. On the

other hand, feature selection algorithms evaluate the "importance" of a subset of features as a whole. This does not mean that all (or any) of the features in the subset have high individual importance. In the context of biomarker discovery, the task of feature selection would be more appropriate for diagnostic markers while feature ranking would be more useful when searching for individual drug targets.

The estimation of importance in feature selection and feature ranking is different. In feature selection, the feature subsets are evaluated explicitly via a predictive model (classifier), built from just those features. As for feature ranking, there is no direct way of evaluating the "correctness" of the order of the individual features. Therefore, our work in this paper focuses on developing an evaluation methodology for feature rankings.

We present our work as follows: First, in Section 2 we define the problem under consideration. We then propose and describe our evaluation methodology in Section 3, where we also consider different approaches of aggregating feature rankings. In Section 4 we outline the experimental evaluation and provide description of the data used. The outcome of the experiments is presented in Section 5. Finally, we discuss the results and draw some conclusions in Section 6.

2 Problem description

We formalize the problem setting as follows: given is dataset D, consisting of k instances (samples) $D = \{S_1, S_2, ..., S_k\}$. Each sample is a vector of n values, $S_i = (v_{i1}, v_{i2}, ...v_{in})$. Each value of an instance represents a certain property or a so-called feature f of that instance. Each feature has a specific value for a specific sample, i.e., $f_j(S_i) = v_{ij}$. Simply put, each row in a dataset is an instance S_i, and each column is the vector of values of a feature f_j.

In this kind of a setting, a feature of particular interest is called a target feature f_{target}, for example the status of some disease. If we apply on the dataset D a ranking algorithm $R(D, f_{target})$, it outputs a list of features $F = [f_1, ..., f_n]$, ordered by decreasing importance $Imp(f_j)$ with respect to f_{target}. The function $Imp(f_j)$ is different for different ranking methods.

In this paper we would like to evaluate how correct is the ordering of features in the ranked list, considering that we never know the ground truth ranking. We will refer to this problem as a problem of evaluating *feature rankings*. This kind of an evaluation methodology, in terms of biomarker discovery, would help answer the question: Which ranking method and/or which study, produce the most "correct" ranked list of genes?

3 Methodology

3.1 Error curve

We approach the problem of evaluating feature rankings by following the idea that the "correctness" of the feature rank is related to predictive accuracy. A

good ranking algorithm would put on top of a list a feature that is most important, and at the bottom a feature that is least important w.r.t. some target concept. All the other features would be in-between, ordered by decreasing importance. By following this intuition, we evaluate the ranking by performing a stepwise feature subset evaluation, with which we generate a so-called *error curve*.

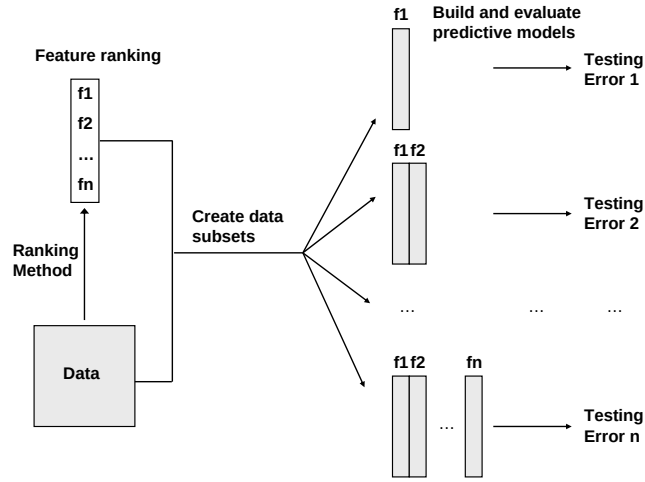


Fig. 1. Constructing an error curve

We present the process of generating the error curve on Figure 1. We begin with a dataset D on which we apply an arbitrary ranking algorithm R. This produces a feature ranking $F = [f_1, ..., f_n]$, where f_1 denotes the top-ranking feature and f_n the bottom-ranked one. We then proceed by generating n data subsets $\{D_{f_{1..1}}, D_{f_{1..2}}..., D_{f_{1..n}}\}$ from the original dataset D. We construct the first data subset $D_{f_{1..1}}$ with only the top-ranked feature f_1. We then add to this subset the second ranked feature f_2, denoted by $D_{f_{1..2}}$. This process is continued iteratively until we add the bottom ranked feature f_n to the $D_{f_{1..n-1}}$ subset, thus yielding $D_{f_{1..n}}$. Finally, we build n predictive models from each of the data subsets and we estimate their error. The points of the error curve are each of the n estimated errors $[E_1, ..., E_n]$. This procedure is summarized in Table 1.

The main idea behind the experimental setup is to evaluate and compare the behavior of different ranking algorithms and different aggregation methods, on single studies and as well across studies.

3.2 Aggregating feature rankings

We consider aggregating feature rankings an important practical issue when working with high-dimensional data. Considering the plethora of feature ranking

Table 1. Constructing an error curve

Input: Data D, Ranking method R
Output: Error curve E
$E \Leftarrow \emptyset$
$D_{f_{1..0}} \Leftarrow \emptyset$
$F \Leftarrow FeatureRanking(R, D)$
for $i = 1$ to n **do**
$D_{f_{1..i}} \Leftarrow D_{f_{1..i-1}} \cup f_i$
$P_i \Leftarrow BuildPredictiveModel(D_{f_{1..i}})$
$E \Leftarrow E \cup EstimateError(P_i)$
end for
return E

methods and datasets that are available, it is reasonable to assume that it might be beneficial to join the different information (rankings) that they provide.

When aggregating feature rankings, there are two issues to consider. The first one is which base feature rankings to aggregate. There are different ways to generate the base feature rankings: from the same dataset, but by different ranking method; from different datasets but the same ranking method or from different subsamples of the same dataset and the same ranking method. The second issue concerns the type of aggregation function to use. Many functions available, and we believe that this is a topic worth exploring by itself, which is out of the scope of this paper. For our initial experiments we used simple methods, like taking the mean or median of the ranks.

4 Experimental setup

4.1 Data description

We performed our experiments on Neuroblastoma studies. Neuroblastoma is the most common extracranial solid tumor of childhood. We considered the status of relapse/no relapse of a patient, as a target concept of interest. The derived markers could be useful for determining the course of treatment of a patient.

We focus on three Affymetrix microarray studies, namely: DePreter et al. [3] (17 samples), Schramm et al. [11] (63 samples) and Wang et al. [12] (99 samples). For practical purposes when presenting the results we refer to them as the "D", "S" and the "W" study.

4.2 Experimental design

We can divide our experiments in two parts: individual study evaluation and cross-study evaluation.

In the individual study setting, we focus on comparing the performance of different ranking approaches. We considered four different feature ranking methods: a simple method based on Information Gain and also more complex methods like Random Forests [2], the ReliefF algorithm [9] and SVM [5]. Furthermore, we investigate if it is beneficial to aggregate the feature rankings produced by different methods on the same study, intuitively similar to [10] and [6]. We considered simple aggregation methods as the Mean rank, Median rank, as well as Min and Max rank.

When investigating the cross-study setting, we considered only one ranking method, namely ReliefF. The idea initially is to compare how feature rankings learned on one study behave if they are tested on another study. Then we examine how that compares to aggregating feature rankings from two different studies and testing on the third.

Table 2. Cross-study evaluation

$S \Rightarrow D$	$D \Rightarrow S$	$D \Rightarrow W$
$W \Rightarrow D$	$W \Rightarrow S$	$S \Rightarrow W$
$agg\{S, W\} \Rightarrow$ D	$agg\{D, W\} \Rightarrow$ S	$agg\{D, S\} \Rightarrow$ W

We summarize the cross-study setting in Table 2. We use "D","S" and "W" to denote different studies and "$A \Rightarrow B$" to signify that we build the feature ranking on study "A" and evaluate it on study "B". When aggregating the feature rankings from the different studies ($agg\{...\}$), we used the previously mentioned aggregation methods.

In both experimental settings, for estimating the error we used the .632+ Bootstrap method [4]. This method combines the leave-one-out cross validation with bootstrap re-sampling. As noted in [4] and [1] this method is well suited for error estimation, especially when working with high-dimensional data. In our setting, we use 20 bags (bootstrap re-sampling), which was previously empirically estimated. We use Naive Bayes as a predictor when constructing the error curve. Although other predictive models were also tested (e.g. SVMs, Decision Trees), we chose Naive Bayes as a method that is dissimilar to any of the feature ranking methods used. For example, we did not wanted to use SVM-RFE as a feature ranking method and then use SVM as a predictive model for evaluation, in order to avoid favoring SVMs as a ranking method.

5 Results

5.1 Individual studies

We present the testing error curves from the single study experiments on Figure 2. On the left-hand side, we show the comparison between the different ranking algorithms, while on the right-hand side the error curves of different

aggregation methods are shown. The figures are ordered in such a way that the results for the smallest dataset (De Preter) are the first figures in a column, while for the largest one (Wang) the results are the last ones in a column.

If we first consider the comparison of different ranking algorithms, it is not immediately obvious which one performs the best. However, it seems that SVM-RFE and ReliefF seem to produce the best ranking, according to the error curves. Also, there is a noticeable effect of the dataset size, where the biggest difference in the curves is for the smallest (De Preter) dataset. Furthermore, if we take a look at the comparisons between the different ranking aggregation methods, the median method has an overall "better" error curve. The median error curve is comparable to the individual ranking algorithms, but it is noticeably less variable.

5.2 Cross studies

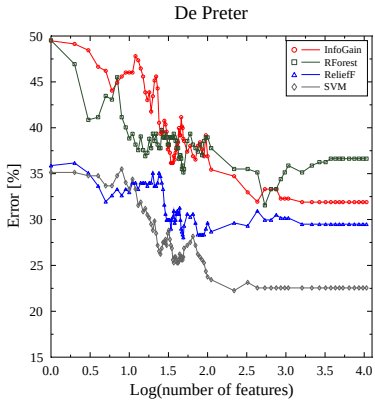
In a similar fashion, we present the results from the cross-studies experiments on Figure 3. The results from the different aggregation methods that are used for combining the feature rankings from the different studies are on the right-hand side figures ((b), (d) and (f)). The comparison between the single study feature ranking and the best aggregated feature ranking, tested on a separate study, are presented on the left-hand side ((a), (c) and (e)). The ordering according to dataset size, also applies here.

The comparison between the different aggregation methods, does not reveal a noticeable difference, although when testing on smaller studies there is great variability of the error curves as compared to testing on bigger studies.

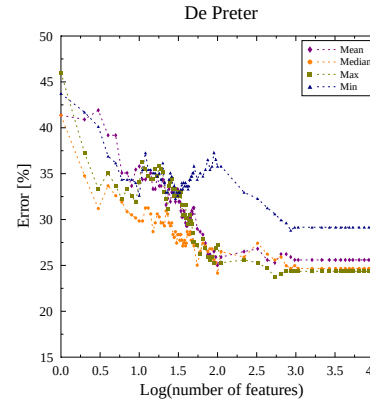
If we take a look at Figure 3(a), it compares between three different feature rankings tested on the De Preter dataset. The feature ranking from the biggest dataset (Wang) is better, but it is worse than the feature ranking produced by aggregating the two different rankings from the Schramm and Wang datasets.

When testing on the Schramm dataset (3(c)), the feature ranking from the smallest dataset (De Preter), performs obviously much worse than the one derived from the biggest dataset (Wang). However, aggregating the feature rankings also does not produce a better ranking. We believe that this is due to the fact that when combining the feature rankings from the two studies, the De Preter derived one is of much worse quality and therefore it has a detrimental effect on the overall aggregated rank.

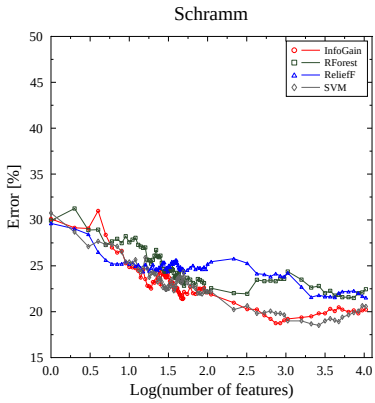
Finally, we show the error curves, when testing on the Wang dataset (Figure 3(e)). On first look, the error curve of the feature ranking derived from the aggregation, seems to be somewhat better than the others. Although a little after the beginning of the curves the error seems to be the same, the curve from the aggregated feature rankings is much less variable than the others. Also it seems that at a very later stage it improves, which we think is due to aggregating an unreliable feature ranking derived from a particularly small dataset.



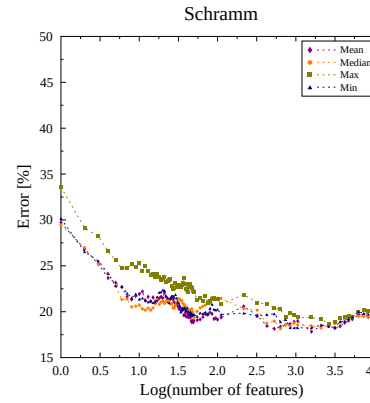
(a) Different ranking methods



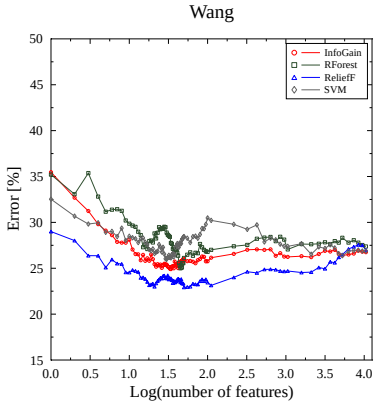
(b) Different aggregation methods



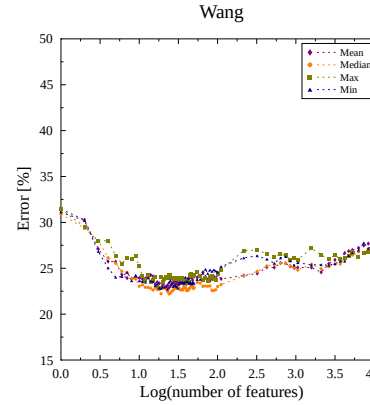
(c) Different ranking methods



(d) Different aggregation methods

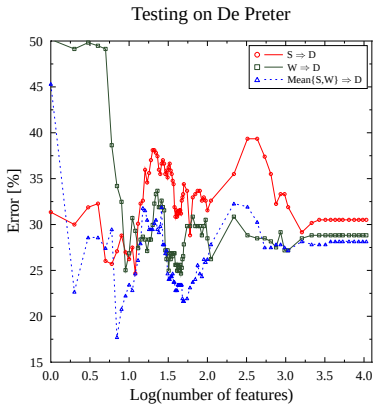


(e) Different ranking methods

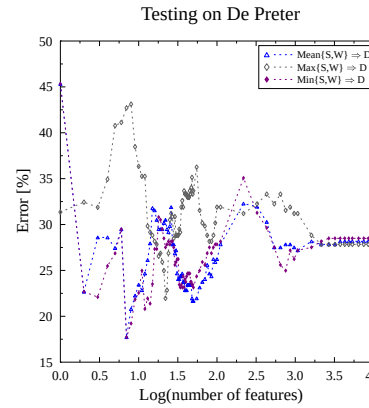


(f) Different aggregation methods

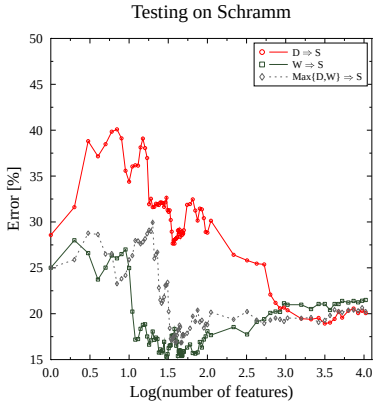
Fig. 2. Single study comparisons



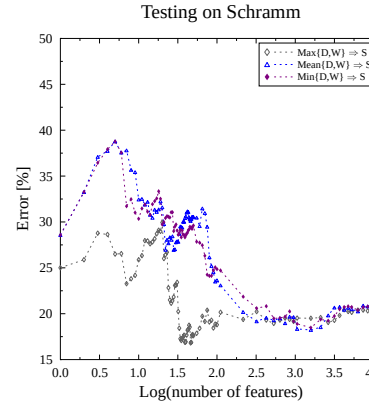
(a) Single vs. combined studies



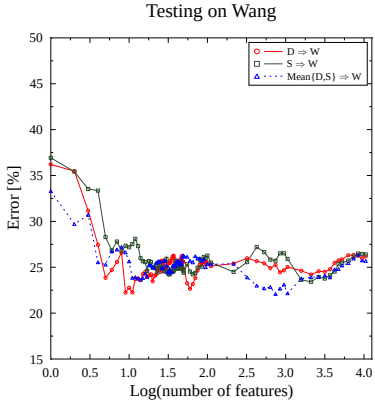
(b) Different aggregation methods



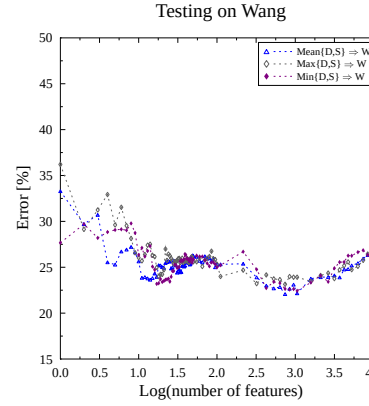
(c) Single vs. combined studies



(d) Different aggregation methods



(e) Single vs. combined studies



(f) Different aggregation methods

Fig. 3. Cross study comparisons

6 Conclusions and further work

In this paper we presented a methodology for evaluating feature rankings. The method relates the "correctness" of the feature ranking to the notion of error of predictive models. We use the so-called error curve, constructed as described in Section 3.1, as an indicator for the quality of the produced feature rankings.

Furthermore, the developed method is used for comparing different ranking approaches and different aggregation approaches for combining feature rankings. From the results presented in Section 5 we can discern two interesting points.The first is related to the size of the error of the curves and the second is related to the variability of the error curves.

Concerning the error size, it is difficult to say with certainty which one is the best feature ranking method or aggregation approach. However, for the ranking methods, it seems that ReliefF and SVMs have the lowest errors. When aggregating feature rankings from different methods, the median aggregation function seems to have the lowest error. The differences in error are very much related to the study size, where bigger differences between ranking algorithms appear for smaller dataset sizes.

The aggregation function used when aggregating feature rankings from different studies seems not to have a particular effect on the testing error. However, when comparing the error curves of feature rankings produced by a single study and the aggregated ones, there is an obvious decrease in the error size. This is especially visible when combining bigger with smaller datasets, although sometimes a too small dataset might have detrimental effect on the aggregated ranking. This is very intuitive, and as a part of our further work we plan to take this into account when performing the aggregation by putting different weights of the base feature rankings related to dataset size and ranking quality.

Another important aspect of the error curve is its variability. One general pattern which can be noticed is that when aggregation of the feature rankings is performed (multiple ranking algorithms or multiple studies), the curve is much less variable than the curves of the base feature rankings. Although the variability does not directly represent feature ranking stability as described in [7] and [8], we believe that it is indicative of it.

In our further work we plan to go beyond the visual inspection of the error curves. The first step would be to use the "area under the error curve" as a numerical way of assessing the quality of the curves. Also, we plan to include a correlation based indicator of stability of the feature rankings, which combined with the area under the curve would provide an insight into the overall quality of the feature ranking.

Acknowledgment

This work was supported by the FP6, E.E.T.-Pipeline project, under the contract LifeSciHealth-2005-037260.

References

1. C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*, 99(10):6562–6566, May 2002.
2. L. Breiman. Random forests. *Machine Learning*, V45(1):5–32, October 2001.
3. K. De Preter, J. Vandesompele, P. Heimann, N. Yigit, S. Beckman, A. Schramm, A. Eggert, R. L. Stallings, Y. Benoit, M. Renard, A. De Paepe, G. Laureys, S. Påhlman, and F. Speleman. Correction: Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes. *Genome Biology*, 8:401+, January 2007.
4. B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.
5. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
6. K. Jong, J. Mary, A. Cornuéjols, E. Marchiori, and M. Sebag. Ensemble feature ranking. In *Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Lecture Notes In Computer Science; Vol. 3202, pages 267–278, 2004.
7. G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, 24(2):258–264, January 2008.
8. A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1):95–116, May 2007.
9. I. Kononenko. Estimating attributes: Analysis and extensions of relief, 1994.
10. Y. Saeys, T. Abeel, and Y. Van de Peer. Robust feature selection using ensemble feature selection techniques. In *ECML/PKDD (2)*, pages 313–325, 2008.
11. A. Schramm, J. H. Schulte, L. Klein-Hitpass, W. Havers, H. Sieverts, B. Berwanger, H. Christiansen, P. Warnat, B. Brors, J. Eils, R. Eils, and A. Eggert. Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene*, aop(current), 2005.
12. Q. Wang, S. Diskin, E. Rappaport, E. Attiyeh, Y. Mosse, D. Shue, E. Seiser, J. Jagannathan, S. Shusterman, M. Bansal, D. Khazi, C. Winter, E. Okawa, G. Grant, A. Cnaan, H. Zhao, N. Cheung, W. Gerald, W. London, K. K. Matthay, G. M. Brodeur, and J. M. Maris. Integrative Genomics Identifies Distinct Molecular Classes of Neuroblastoma and Shows That Multiple Genes Are Targeted by Regional Alterations in DNA Copy Number. *Cancer Res*, 66(12):6050–6062, 2006.

A Subgroup Discovery Approach for Relating Chemical Structure and Phenotype Data in Chemical Genomics

Lan Umek[1], Petra Kaferle[2], Mojca Mattiazzi[2], Aleš Erjavec[1], Črtomir Gorup[1], Tomaž Curk[1], Uroš Petrovič[2] and Blaž Zupan[1,3]

[1] Faculty of Computer and Information Sciences, University of Ljubljana, Slovenia
[2] Jožef Stefan Institute, Ljubljana, Slovenia
[3] Dept. of Human and Mol. Genetics, Baylor College of Medicine, Houston, USA

Abstract. We report on development of an algorithm that can infer relations between the chemical structure and biochemical pathways from mutant-based growth fitness characterizations of small molecules. Identification of such relations is very important in drug discovery and development from the perspective of argument-based selection of candidate molecules in target-specific screenings, and early exclusion of substances with highly probable undesired side-effects. The algorithm uses a combination of unsupervised and supervised machine learning techniques, and besides experimental fitness data uses knowledge on gene subgroups (pathways), structural descriptions of chemicals, and MeSH term-based chemical and pharmacological annotations. We demonstrate the utility of the proposed approach in the analysis of a genome-wide *S. cerevisiae* chemogenomics assay by Hillenmeyer *et al.* (Science, 2008).

1 Introduction

One of the promises of the post-genomics era was the identification of novel drug targets and design of more efficient and specific drugs with fewer side-effects. In reality, pipelines of pharmaceutical companies did not improve much due to genomic data alone. One of the main reasons for that is the lack of methods to combine characteristics of potential drug molecules with rich genomic data. Such data comes in many flavors: from raw genome sequence data, phenotypic profiles such as gene expression profiles, functional and physical interactions of genes and proteins, to rich annotations of genes and their products by complex ontologies. These together define phenomes (*i.e.*, genome-wide phenotypes) of a cell or an organism.

Of special interest for the identification and characterization of potential drug molecules are recently developed chemogenomic approaches. These profiles allow to measure changes in the phenome that were caused by the molecule's activity. When applied to a collection of mutants, we gain a data set with a vast potential for the generation of chemogenomics hypotheses. One such data set was recently reported by Hillenmeyer *et al.* [1], where growth fitness in the presence of a number of chemicals was observed in a set of genome-wide single-gene

deletion mutants. The authors used yeast *S. cerevisiae* as a model organism, and reported that a surprisingly large proportion (97%) of gene deletions exhibited a measurable growth phenotype.

The aim of the research reported here was to see if the data set published by Hillenmeyer *et al.* [1] could be used further to relate genetic pathways with structural and pharmacological properties of drugs. We extended the information from fitness data by associating chemicals with their structural descriptions, and mined subsets of mutants that stem from single-deletions of genes common to a specific pathway. Our effort, in which we queried a number of data bases to complement experimental results and to allow for further analysis, could be enlisted under *integrative bioinformatics* or *integrative systems biology*. These emerging fields strive to relate a plethora of existing molecular biology data bases and experimental repositories [2].

To serve our aim, we developed a specific data mining approach. In particular, we used a combination of unsupervised learning (clustering) to find groups of chemicals with similar mutant-based fitness profiles, and supervised learning to check if discovered groups of chemicals can be characterized in terms of common chemical structure. The proposed search algorithm evaluates such hypotheses across a number of genetic pathways and tests a variety of plausible subgroups of chemicals. While the particular approach is new and for the first described in this report, it in part resembles rule-based subgroup discovery techniques [3, 4] and bi-clustering approaches [5]. From the former, we borrow the idea of finding subsets of characteristic data items, and from the latter the idea that items have to be similar in two different aspects, in our case, in structure of the chemicals and corresponding phenotype response.

The paper proceeds with the description of the data set used in our experiments, and of preprocessing (data selection) steps. We continue with a detailed description of the algorithm, experimental results and a discussion.

2 Data

In early 2008, Hillenmeyer *et al.* published a comprehensive analysis that included 1144 chemical genomic assays on the yeast whole-genome heterozygous and homozygous gene deletion collections and quantified the growth fitness of each deletion strain in the presence of chemical or environmental stress conditions [1]. This study generated the first available data set based on which systematic analysis of functional relations between biochemical pathways and chemical structure is possible.

In the analysis reported here we focussed on experiments on homozygous strains. From the initial set of 418 genome-wide screens, we removed experiments with environmental stress, irradiated drugs, inorganic compounds, platinum compounds, norcantharidin, cantharidin analog and cantharidin disodium, with the aim to focus on the chemical space of organic substances. We also discarded assays for which the molecular formula was unavailable according to the

supplementary data [1], assays that used mixtures of two chemicals and experiments where time of growth was different than 20 generations.

From the remaining 136 assays, the filtering of assays with the same small molecule at different concentrations was based on manual inspection of graphs of quantile functions of fitness values (Figure 1). From such sets of experiments, we selected the one with the sharpers transition in the related graph. In almost all cases the lowest concentration was selected. If two assays used the same chemical at the same concentration, only the first assay listed in the data was used. The resulting set included 74 assays.

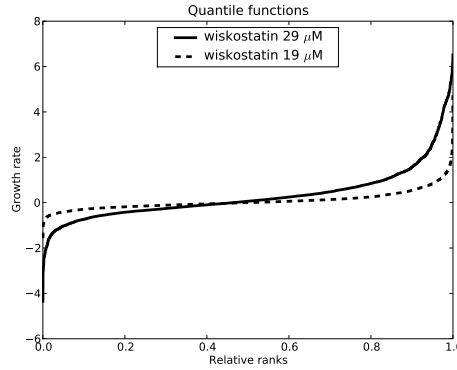


Fig. 1. Quantile function of mutant growth fitness for wiskostatin is shown. Here, the experiment concentration of $19\mu M$ was selected due to sharper quantile function.

We used NCBI's PubChem (`http://pubchem.ncbi.nlm.nih.gov/`) to obtain SMILES structural descriptors [6]. We failed to do so for three chemicals in the collection, and thus proceeded with 71 chemicals and their related assays. SMILES descriptors were converted to an array of molecular descriptions (constitutional and topological descriptors, molecular properties, connectivity indices, atom-centred fragments, functional group counts) using the Dragon [7] software. We removed constant, near-constant, and highly correlated (correlation exceeded 0.9) descriptors and drug-like indices (like Ghose-Viswanadhan-Wendoloski index). For the analysis, we used 126 molecular descriptors.

The resulting 71 assays, each corresponding to an application of a specific small molecule, included growth fitness measurement of 4262 single-mutants. Of these, we have removed 507 mutants with missing growth fitness values. The above preprocessing thus produced a data matrix that included the growth fitness scores for 3755 single gene deletion mutants in the presence of 71 different assays.

3 Methods

We will assume that our data is a random sample $e_1 = (X_1, Y_1), \ldots, e_n = (X_n, Y_n)$, $(n = 71)$ where each pair $e_i = (X_i, Y_i)$ represents a chemogenomical experiment. Each experiment consists of two vectors: X_i is a set of DRAGON-based structural descriptors of the i-th chemical used in the assay, and Y_i is a resulting vector of phenotype responses, consisting of growth fitness scores for 3755 single gene deletion budding yeast mutant.

We here propose a method that aims to relate chemical structures of the small molecules involved in experiments with their characteristic phenotypic profile. In particular, we are looking for subgroups of experiments (chemicals) where:

- experiments in the subgroup have similar phenotypic profile in some specific subsets of mutants,
- the set of chemicals in the subgroup can be reliably discriminated from other chemicals in the data set using DRAGON-based structural descriptions.

The subsets of mutants were identified based on the annotation of a gene to a specific KEGG pathway [8]. We have only used pathways that include more than two mutated genes. As of April 2009, there were 98 such pathways, covering 760 genes in total.

We have applied a specific search algorithm that uses unsupervised learning to find subgroups of chemicals with similar gene set-based phenotypic profile, and supervised learning to identify those subgroups which can be successfully characterized by the set of chemical structure descriptors. The final step of the analysis is a MeSH term enrichment-based characterization of resulting subsets of chemicals. Both steps, the search algorithm and chemical characterization of the subsets are described below.

3.1 Search Algorithm

The algorithm searches for characterizable sets of chemicals that resulted in similar phenotypic profiles for a subset of mutants. The algorithm is executed all gene sets (one KEGG pathway represents one gene set), and includes the following steps:

1. Choose a subset of phenotypic features (genes from a specific KEGG pathway) GS and define the dissimilarity measure δ_{GS} between two experiments e_i, e_j (phenotypic profiles) using a weighted Manhattan metric:

$$\delta_{GS}(e_i, e_j) = \sum_{k \in GS} \frac{|Y_{ik} - Y_{jk}|}{\max_l Y_{lk} - \min_l Y_{lk}} \tag{1}$$

   where Y_{ik} represents k–th component of random vector Y_i.
2. Perform hierarchical clustering [9] of the experiments with δ_{GS} using Ward's minimum-variance linkage [10].

3. Traverse the resulting dendrogram to identify various candidates for subgroups. Consider only subgroups consisting of at least $min_{size} = 4$ chemicals and no bigger than $max_{size} = 10$.
4. For all subgroups (of chemicals) identified in the previous step, estimate the degree of separability from the rest of the chemicals in the data set. For each subgroup, we first classify chemicals based on their membership in the subgroup. We then perform leave-one-out to estimate the accuracy of support vector machine (SVM)-based class-prediction. SVM is presented with DRAGON-based chemical structure descriptors and the classification in the current subgroup. Area under ROC (AUC) is used to measure the predictive accuracy. SVM with linear kernel as implemented in SVMlight library (Linear Learner with default parameters) [11] was used in our experiments. Subgroups with AUC equal to 0.75 or above are retained and reported to the user.

3.2 Characterization of Subgroups

The discovered subgroups include a set of chemicals which share a similar phenotype response in a KEGG pathway-specific subset of mutated genes. Each reported subset of mutated genes is therefore characterized by the name of their respective KEGG pathway. We also need a simple, readable characterization of chemicals in the subgroup. For this, we have used terms from the *chemical classification* and *pharmacological classification* part of Medical Subject Headings (MeSH) ontology. Annotations of chemicals with MeSH terms were retrieved from NCBI's PubChem (`http://pubchem.ncbi.nlm.nih.gov/`). We then used enrichment analysis to find terms characteristic for a given subset of chemicals. Given all chemicals annotated to term t and chemicals in subgroup G we test if there exists a relationship between membership of the subgroup G and term t using *Fisher's exact test*. The p–values from Fisher's test are then used for ranking annotated terms. We report terms with the associated p–value less than 0.05.

3.3 Implementation

The proposed method was developed in Python within the Orange data mining framework [12], which implements unsupervised and supervised techniques, leave-one-out evaluation and ROC analysis. Orange Bioinformatics toolbox [13] was used to access KEGG pathways, obtain MeSH terms and chemical annotations, and perform enrichment analysis of MeSH terms.

4 Results

Our algorithm discovered 25 subgroups. Eleven of them resulted in at least one enriched classification term (pharmacological or chemical), eight of them (for which all terms were annotated to at least 2 small molecules) are presented in

Table 1. They include 40 small molecules (56.5% of the experiments). The highest AUC score was 0.876 for a subgroup not shown in the Table (no associated enriched terms).

Table 1. A selection of subgroups (chemicals and their associated phentoypic profiles) as discovered by the proposed algorithm. Reported are the number of small molecules in a subgroup, AUC scores, associated KEGG pathway, and enriched chemical and pharmacological MeSH terms.

size	AUC	pathway	chemical classification	pharmacological classification
5	0.855	nitrogen metabolism	sulfur compounds	myeloablative agonists
				toxic actions
5	0.855	ubiquinone biosynthesis	hydrocarbons, halogenated,	antineoplastic agents
			nitrogen mustard compounds	alkylating
7	0.819	biosynthesis of steroids	disulfides	none
5	0.782	drug metabolism		
		other enzymes	urea	none
5	0.779	alanine and aspartate		
		metabolism	disulfides	none
6	0.756	cell cycle - yeast	disulfides, allyl compounds	protective,
				anticarcinogenic agents
6	0.756	folate biosynthesis	azirines, sulfur compounds	antineoplastic,
				alkylating agents
8	0.752	one carbon pool	allyl compounds	protective, antineoplastic,
		by folate		anticarcinogenic agents

5 Discussion

The algorithm presented in this paper enables inference relations between chemical structure and biochemical pathways. Identification of such relations is very important for drug discovery, since it allows for an argument-based selection of candidate molecules in target-specific screenings, and early exclusion of substances with highly probable undesired side-effects.

The experimental analysis we report in the paper uses the first, and currently the only publicly available data set that observes chemically-induced phenotypes in a genome-wide set of single-gene mutations. With availability of single-mutant collections for a range of model organisms, and promises of corresponding RNA-interference platforms that could also be applied for genome-wide phenotype screening of human samples, we expect the emergence of similar data sets in the near future. The presented computational approach should therefore not be regarded as a single-application attempt, but rather as an enabling technology that could help us in data analysis and hypothesis formation from the soon-to-emerge experimental data.

The comprehensive evaluation of the results in the Table 1 is beyond the scope of this paper. An ultimate test would require a number of wet-lab exper-

iments to either confirm or discard the proposed hypotheses. We have, though at the scanning stage, found some of the proposed hypotheses very interesting. One of the identified subgroups, consisting of six molecules (Figure 2), is related to the cellular process "cell cycle". Disturbances in cell cycle regulation are the hallmark of cancer. Enrichment analysis of the chemicals revealed that the subgroup contains both anticarcinogenic agents from the data set, parthenolide and amsacrine. Moreover, the most characteristic phenotypic marker by which the six substances were clustered together was the relative growth fitness of mutants in three genes (*LTE1*, *DBF2* and *CDH1*), which are all involved in mitotic exit. Parthenolide, the more thoroughly studied of the two identified anticarcinogenic substances, is indeed thought to affect this phase of the cell cycle [14]. This example thus illustrates the biological relevance of the proposed method and illustrates usefulness of such methods. For example, identification of anticarcinogenic activity of parthenolide was identified in an experimental screen [15] of the type which is notoriously error-prone. The method presented here demonstrates that such computational analysis prior to experimental screens could importantly increase the likelihood of a positive outcome of the screens.

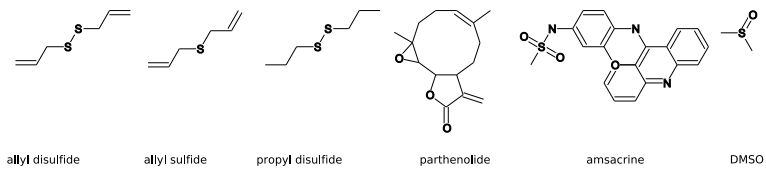


Fig. 2. Subgroup of six small molecules having similar impact on cell cycle genes. Their enriched chemical terms are disulfides and allyl compounds ($p = 0.0048$), their enriched pharmacological terms are anticarcinogenic agents ($p = 0.0055$) and protective agents ($p = 0.0161$).

6 Conclusions

This report presents the first attempt to analyze chemogenomic data by relating chemical structures to biochemical pathways. An example is given to demonstrate the biological relevance of the proposed method. It should be noted, however, that for the full extent of the usefulness of the proposed method, more comprehensive data sets are required. As a lesson from the study, screens uniformly covering the chemical space in the selection of tested molecules are likely to provide the best predictive power. Further impact of the combined experimental and computational methods described here for drug discovery and development will be achieved when technical limitations for conducting genome-wide screens in mammalian cells will be overcome; importantly, the method presented here for yeast data is, with only slight modifications, useful also for mammalian systems.

Acknowledgment

The study was supported by grants from Slovenian Research Agency (J2-9699, L2-1112).

References

[1] Hillenmeyer, M.E., Fung, E., Wildenhain, J., Pierce, S.E., Hoon, S., Lee, W., Proctor, M., St, Tyers, M., Koller, D., Altman, R.B., Davis, R.W., Nislow, C., Giaever, G.: The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. Science **320**(5874) (2008) 362–365

[2] Hoon, S., Smith, Wallace, I.M., Suresh, S., Miranda, M., Fung, E., Proctor, M., Shokat, K.M., Zhang, C., Davis, R.W., Giaever, G., St Onge, R.P., Nislow, C.: An integrated platform of genomic assays reveals small-molecule bioactivities. Nat Chem Biol **4**(8) (2008) 498–506

[3] Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research **5** (2004) 153–188

[4] Ženko, B., Struyf, J.: Learning predictive clustering rules. In: 4th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers, volume 3933 of LNCS, Springer (2005) 234–250

[5] Van Mechelen, I., Bock, H.H., De Boeck, P.: Two-mode clustering methods: a structured overview. Statistical Methods in Medical Research **13**(5) (2004) 363–394

[6] Weininger, D.: SMILES, a chemical language and information system. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences **28**(1) (1988) 31–36

[7] Talete srl: Dragon for Windows (Software for Molecular Description Calculations), Version 5.5 (2007)

[8] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y.: Kegg for linking genomes to life and the environment. Nucleic Acids Research **36**(Database issue) (December 2007) D480–D484

[9] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley (1990)

[10] Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301) (1963) 236–244

[11] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874

[12] Demšar, J., Zupan, B., Leban, G.: Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper (2004)

[13] Curk, T., Demšar, J., Xu, Q., Leban, G., Petrovič, U., Bratko, I., Shaulsky, G., Zupan, B.: Microarray data mining with visual programming. Bioinformatics **21**(3) (2005) 396–8

[14] Fonrose, X., Ausseil, F., Soleilhac, E., Masson, V., David, B., Pouny, I., Cintrat, J.C., Rousseau, B., Barette, C., Massiot, G., Lafanechere, L.: Parthenolide Inhibits Tubulin Carboxypeptidase Activity. Cancer Res **67**(7) (2007) 3371–3378

[15] Jonathan J. Ross, J.T.A., Birnboim, H.C.: Low Concentrations of the Feverfew Component Parthenolide Inhibit In Vitro Growth of Tumor Lines in a Cytostatic Fashion. Planta Med **65**(2) (1999) 126–129

Part III

Poster Abstracts

Robust biomarker identification for cancer diagnosis using ensemble feature selection methods

Thomas Abeel[1,2], Thibault Helleputte[3,4],Yves Van de Peer[1,2], Pierre Dupont[3,4], and Yvan Saeys[1,2]

[1] Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium
[2] Department of Molecular Genetics, Ghent University, Gent, Belgium
{thomas.abeel, yves.vandepeer, yvan.saeys}@psb.ugent.be
[3] Department of Computing science and Engineering INGI Université catholique de Louvain
{thibault.hellepute, pierre.dupont}@uclouvain.be

Introduction: Biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high dimensional data. Surprisingly, the stability with respect to sampling variation or robustness of such selection processes has received attention only recently. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. In addition, a more robust set of markers may strengthen the confidence of an expert in the results of a selection method.

Methodology: Our first contribution is a general framework for the analysis of the robustness of a biomarker selection algorithm. Secondly, we conducted a large scale analysis of the recently introduced concept of ensemble feature selection, where multiple feature selections are combined in order to increase the robustness of the final set of selected features. We focus on selection methods that are embedded in the estimation of support vector machines (SVMs). SVMs are powerful classification models that have shown state-of-the-art performance on several diagnosis and prognosis tasks on biological data. Their feature selection extensions also offered good results for gene selection tasks.

Results: We show that the robustness of SVMs for biomarker discovery can be substantially increased by using ensemble feature selection techniques, while keeping the same classification performances. The proposed methodology is evaluated on four microarray data sets showing increases of up to 27% in robustness of the selected biomarkers. The stability gain obtained with ensemble methods is particularly noticeable for small signature sizes (a few tens of genes), which is most relevant for the design of a diagnosis or prognosis model from a gene signature.

Keywords: Feature selection, Support Vector Machines, Biomarker discovery

Java-ML a Java Library for Data Mining

Thomas Abeel, Yves Van de Peer, and Yvan Saeys

Department of Plant Systems Biology, VIB, Ghent University, Gent, Belgium
{thabe, yvpee, yvsae}@vib-ugent.be

Introduction: In this abstract we describe Java-ML [1]. Java-ML is a collection of machine learning algorithms, which aims to be a readily usable and extensible API for both software developers and research scientists . Several well-known data mining libraries already exist, including Weka and Yale/RapidMiner. In contrast to these tools, Java-ML is oriented towards users who write their own programs. To this end, the interfaces are restricted to the essentials, and are easy to understand.

Description: Java-ML contains an extensive set of similarity-based techniques, and offers state-of-the-art feature selection techniques. The large number of similarity functions allows for a broad set of clustering and instance-based learning techniques. The feature selection techniques are well-suited to deal with high-dimensional domains often encountered in bioinformatics. Using Java-ML in your own software is easy. For example, the following lines of code integrate K-Means clustering into your code.

```
Dataset data = FileHandler.loadDataset(new File("iris.data"), 4, ",");
Clusterer km = new KMeans();
Dataset[ ] clusters=km.cluster(data);
```

The first line loads data from the iris.data file, which has the class label in the fourth column, and the fields are separated by a comma. The second line constructs a new instance of the KMeans clustering algorithm with default values, in this case k=4. In the third line data is clustered and the clusters are returned. There are several sources of documentation for Java-ML: the source code itself is documented, the website provides a number of tutorials with annotated code samples for common tasks and, finally, the website also has the API documentation of all releases.

Conclusion: In this abstract, we described Java-ML, a library of machine learning algorithms, available from `http://java-ml.sf.net/` under the GNU GPL license.

Keywords: machine learning, library, feature selection, instance based learning, clustering

References

1. Thomas Abeel, Yves Van de Peer, and Yvan Saeys. Java-ML: A machine learning library. Journal of Machine Learning Research, 10:931-934, 2009.

Extending KEGG Pathways for a Better Understanding of Prostate Cancer Using Graphical Models

Adel Aloraini[1], James Cussens[1], and Richard Birnie[2]

[1] Artificial Intelligence Group, Computer Science Dept, University of York, YO105DD, UK
{aoraini, jc}@cs.york.ac.uk
[2] Cure Therapeutics Limited Biocentre, Innovation Way Haslington, York, YO105NY, UK
richard.birnie@pro-cure.uk.com

In this ongoing work, partial information about WNT-signaling pathway found in KEGG has been used as prior knowledge to guide a machine learning algorithm, based on linear regression, to show a better understanding of how cellular system works in WNT-signaling pathway. This work is based on a published paper by [1]. We are using a set of graphical models called dependency networks to give a bigger picture of how genes in different components in the pathway affect each other. The ultimate goal is to understand which genes in one component cause which in other components. The heuristic search used in this study is guided by the prior knowledge extracted from WNT-signaling pathway. In this work we have used AIC score function that balance between adding parents to the regression model and the trade-off between bias and variance. Beside AIC score function, for each family in the network, the set of parents are examined by Residual Sum of Squares (RSS). Since if a set of parents in the regression model leads RSS to be zero, this means that AIC will go to infinite value, which in turn gives nonsense result. Therefore, the model from the greedy search that has been used is further examined to see if RSS is non-zero. If RSS equals to zero the correlation coefficient is used to drop parents with small correlation with the child gene .We would emphasize that the resultant network basically looks at which genes from one component causes or react with which in another component. In the future work we will examine the overfitting problem based on cross-validation. After obtaining satisfied result, we will look at the interaction between genes inside each component.

Keywords: WNT-signaling pathway,dependency Networks, heuristic search, correlation coefficient

References

1. Richard Birnie, Steven D Bryce, Claire Roome, Vincent Dussupt, Alastair Droop, Shona H Lang, Paul A Berry, Catherine F Hyde, John L Lewis, Michael J Stower, Norman J.Maitland and Anne T Collins : Gene expression

profiling of human prostate cancer stem cells reveals a pro-inflammatory phenotype and the importance of extracellular matrix interactions. J. Genome Biol 9(5), R83 (2008).

Variable Pruning in Bayesian Sequential Study Design

P. Antal[1], G. Hajós[1], A. Millinghoffer[1], G. Hullám[1], Cs. Szalai[2], and A. Falus[3]

[1] Dept. of Meas. and Inf. Sys., Budapest University of Technology and Economics
[2] Inflammation Bio. and Immunogenomics Research Group, Hungarian Acad. of Sci.
[3] Dept. of Genetics, Cell- and Immunobiology, Semmelweis University, Hungary
antal@mit.bme.hu

The relative scarcity of the results of genetic association studies (GAS) prompted many research directions and hypotheses. Genome-wide association studies (GWAS) had appeared, now followed by whole-genome sequencing and deep sequencing. Based on the complex data the multiple testing problem computer intensive statistical methods became widespread. Despite the proven repeatability and transferability of GWAS results, it is a widely shared belief that the efects of the variations in the human genome, particularly significant for personalized diagnosis are still largely unexplored. The "rare haplotypes" hypothesis exploiting the new generation sequencing techniques targets new variants, whereas the "common disease common variants" hypothesis focuses on the probabilistic interaction and causal relation of many weak factors including environmental efects. We attempt to discover from subsequent measurements of well-selected blocks of single-nucleotide polymorphisms (SNPs) the relevant genetic factors for a given target set, which keeps only the promising variables. It uses interim analysis and meta-analysis of the available aggregated data sets for guiding further measurements. We apply complex Bayesian network based structural features in the analysis, specifically Markov Blanket Memberships (MBM), Markov Blanket Sets (MBS), and Markov Blanket Graphs (MBGs) [1]. The applied Bayesian multilevel relevance analysis means a multivariate approach to discover relevant sets of variables together with their interactions, in correspondence we work with multivariate preferences (utilities). We evaluate typical policies in association studies based on interim Bayesian meta-analysis, and also the performance of a one-step look ahead approximation of the expected value of experiments in the full Bayesian approach. Our application domain is the investigation of genetic background of asthma.

Keywords: adaptive study design, bayesian decision support, bayesian networks

Acknowledgments Supported by grants from the OTKA National Scientifc Research Fund (PD-76348); NKTH TECH-08-A1/2-2008-0120 (Genagrid), and the Jnos Bolyai Research Scholarship of the Hungarian Academy of Sciences (P.Antal).

References

1. P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus. A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. JMLR Proceeding, 4:74-89, 2008.

On the Bayesian applicability of graphical models in genome-wide association studies

P. Antal[1], A. Millinghoffer[1], Cs. Szalai[2], and A. Falus[3]

[1] Dept. of Meas. and Inf. Sys., Budapest University of Technology and Economics
[2] Inflammation Bio. and Immunogenomics Research Group, Hungarian Acad. of Sci.
[3] Dept. of Genetics, Cell- and Immunobiology, Semmelweis University, Hungary
antal@mit.bme.hu

Probabilistic graphical models are widely applied tools in expression data analysis, in pedigree analysis, in linkage and association analysis (e.g., see [4, 2]). We proposed the use of Bayesian networks in partial genetic association studies as a tool, which can learn non-transitive, multivariate, non-linear relations between target and explanatory variables, treat multiple targets, and allow scalable multivariate analysis [1]. To cope with high sample complexity we used the Bayesian statistical framework, which allows model-averaging as an automated solution for the multiple testing problem and marginalization to the relevant aspects of the model. However the applicability of graphical models in genome-wide association studies is hindered by the high sample size and computational complexity. We present results about the learning characteristics of their Bayesian application, and overview the computational complexity to indicate the achievable gains of parallelization [3]. Specifically, we examine the following.

1. How the sample size affects the posterior of specific feature values (especially regarding the Markov-blanket membership feature).
2. Can an internal score be defined (e.g. the entropy of the posterior) to characterize the applicability of a high-dimensional, multivariate analysis.
3. How sample size affects performance using an external, reference from the point of view of feature subset selection FSS using sensitivity and specificity, misclassification rate and AUC (cf. the multiple testing problem).

Keywords: Bayesian learning, bayesian networks, learning rate

References

1. P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus. A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. JMLR Proceeding, 4:74-89, 2008
2. D. J. Balding. Handbook of Statistical Genetics. Wiley & Sons, 2007.
3. H. J. Cordell. Detecting gene-gene interactions that underlie human diseases. Nature Reviews: Genetics, 10(1):392-404, 2009.
4. N. Friedman. Inferring cellular networks using probabilistic graphical models. Science, 303(5659):799 - 805, 2004.

Averaging over measurement and haplotype uncertainty using probabilistic genotype data

P. Antal[1], P. Sárközy[1], B. Zoltán[1], P. Kiszel[3], Á. Semsei[3], Cs. Szalai[2], and A. Falus[3]

[1] Dept. of Meas. and Inf. Sys., Budapest University of Technology and Economics
[2] Inflammation Bio. and Immunogenomics Research Group, Hungarian Acad. of Sci.
[3] Dept. of Genetics, Cell- and Immunobiology, Semmelweis University, Hungary
antal@mit.bme.hu

The relative accuracy and symbolic nature of the genotype data hide the complexity of preprocessing such as image processing and clustering. Whereas in expression data analysis noise models are common to cope with similar normalization and feature extraction problems, such models are missing in genetic association studies (GAS). We present a probabilistic approach to model uncertainties in genotyping with explicit representation of rejection and a probabilistic averaging framework to cope with such uncertain data. Beside measurement problems, uncertainty also arise in haplotype reconstruction and we present a two-phased Monte Carlo integration of PHASE, an existing haplotype reconstruction method [2] and our earlier Bayesian model-based data analysis [1]. The implemented framework allows the explicit propagation of uncertainty from measurement through haplotype reconstruction to data analysis allowing better understanding of the sufficiency of the measurements.

Keywords: genotyping error,haplotype reconstruction, uncertain data, bayesian feature subset analysis, bayesian networks

Acknowledgments Supported by grants from the OTKA National Scientifc Research Fund (PD-76348); NKTH TECH-08-A1/2-2008-0120 (Genagrid), and the Jnos Bolyai Research Scholarship of the Hungarian Academy of Sciences (P.Antal).

References

1. P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus. A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. JMLR Proceeding, 4:74-89, 2008.
2. M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. The American Society of Human Genetics, 73(5):1162-1169, 2003.

Bayes Meets Boole: Bayesian Learning of Boolean Regulatory Networks from Expression Data

Matthias Böck[1,2], Soichi Ogishima[3], Lars Kaderali[2], and Stefan Kramer[1⋆]

[1] TU München, Institut für Informatik/I12, Boltzmannstr. 3, 85748 Garching b. München, Germany
{matthias.boeck, stefan.kramer}@in.tum.de
[2] University of Heidelberg, Viroquant Research Group Modeling, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany
lars.kaderali@bioquant.uni-heidelberg.de
[3] Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan
ogishima@sysbioevo.org

Modern high-throughput molecular techniques deliver data sets which allow large-scale gene expression analysis and to gain insight into the complex architecture of biological networks. Various reverse engineering methods have been tested so far to deal with the challenge of inferring gene regulatory interactions from these data sets. One main problem is the increasing complexity of defining and calculating detailed parameters and hence computation time. To bypass this issue, we propose a model to infer genetic regulations from gene expression data which uses a discretization to Boolean states to deal with the high complexity of the data. Another assumption of the approach is the use of the sigmoid function, which models dependencies between genes. This is integrated into a probabilistic framework, using a Bayesian approach to infer a network topology from the Boolean gene states. To prevent the algorithm from overfitting noisy data in its prediction, an additional prior is added to regularize the result. The prior is defined in a way that it favors scale-free networks. Prediction accuracy was evaluated on several simulated data sets as well as on a small biological network, the Saccharomyces cerevisiae cell cycle. Additionally, testing was performed on a larger set of 800 cell cycle regulated genes, to test if basic properties of biological networks could be inferred by the algorithm from given gene concentrations. In effect, the algorithm was able to reconstruct from simulated as well as biological data sets main regulatory dependencies of the original topology. It furthermore has a low time complexity and is applicable even to large simulated networks with more than 4000 nodes. Difficulties arise for small networks, where the integrated prior has only little influence and only few measurements are available to describe the interactions. Further tests on a larger set of 800 cell cycle regulated genes revealed that the algorithm can also infer scale-free topologies for large networks, with typical properties of biological networks like robustness or average shortest path lengths.

⋆ Both authors contributed equally.

Keywords: Bayesian learning, Boolean networks, graphical models, gene regulation, regulatory networks, gene expression, systems biology

Statistical relational learning for supervised gene regulatory network inference

Céline Brouard[1], Julie Dubois[1,2], Marie-Anne Debily[3], Christel Vrain[1], and Florence d'Alché-Buc[1]

[1] IBISC CNRS FRE 3190, Universite d'Evry-Val d'Essonne, Evry 91000, FRANCE
{florence.dalche,cbrouard}@ibisc.fr
[2] LIFO, Universite d'Orleans, BP 6759 45067 Orleans Cedex 2, FRANCE
christel.vrain@univ-orleans.fr
[3] CEA, Evry 91000, FRANCE

Starting from a known gene regulatory network involved in the switch proliferation/differentiation in keratinocytes cells, we have developed a new approach to learn rules that can explain the presence or absence of regulation between two genes. For this purpose, we have used experimental data (gene expression) as well as knowledge such as GO annotations and positions of genes on chromosomes. In the context of statistical relational learning, we have learned the concept of transcriptional regulation between two genes, represented by a predicate "regulate" [1], [2], [3]. A network of genes extracted from Ingenuity has been used for labeling couples of genes, and experimental data as well as prior knowledge have been encoded into a first order representation (ground atoms and rules) [2]. We have first applied a pure inductive logic programming approach, Aleph[4] and we have compared it to a statistical relational learning approach [5], called Markov Logic Network, introduced by Domingos et al.[7, 6] In this framework, a set of weighted logical rules is represented by a random Markov network: nodes correspond to ground atoms, rules allow to form cliques, and the weights of the rules are associated to the corresponding cliques. Making a decision corresponds to computing the posterior probability of the labels given the input description. We have used Aleph to produce a large set of rules, thus fixing the structure of the random Markov network, and we have applied a discriminative learning method to get the weights associated to the rules implemented by Alchemy, a source code implemented and described in [8]. Among the rules ranked by Alchemy, we have found interesting regulatory patterns which show that first, Ingenuity can be cross-validated by experimental data and provide consistent information and second, new rules can be used to suggest new candidates for regulators and regulees. However, in terms of performance, Alchemy only marginally improves performance upon Aleph. Results are discussed and compared to those obtained by [9] on other relational tasks.

Keywords: biological network inference, statistical relational learning, gene regulatory network, gene expression

References

1. C. Combe, F. d'Alche-Buc. Apprentissage relationnel du concept de regulation, Eprints PASCAL (2005).
2. J. Dubois. Apprentissage relationnel pour l'inference d'un reseau de regulation, Master Thesis, IMBI, Universite d'Evry, FRANCE (2007).
3. C. Brouard. Apprentissage statistique relationnel pour l'inference d'un reseau de regulation, Master Thesis, Tlecom Bretagne, FRANCE (2007).
4. A. Srinivasan.The Aleph Manual (2004)
5. L. Getoor, N. Friedman, D. Koller, A. Pfeffer, and B. Taskar (2007) Probabilistic Relational Models. In L. Getoor and B. Taskar, editors, Introduction to Statistical Relational Learning.
6. P. Domingos, S. Kok, D. Lowd, H. Poon, M. Richardson, P. Singla. Markov Logic. In L. De Raedt, P. Frasconi, K. Kersting and S. Muggleton (eds.), Probabilistic Inductive Logic Programming, pp. 92-117 (2008)
7. D. Lowd, P. Domingos. Efficient weight learning for Markov logic networks. Proc. of 7th European Conf. of Principles and Practive of Knowledge Discovery in Databases (PKDD-07) 2007.
8. S. Kok, P. Singla, M. Richardson, P. Domingos The Alchemy system for statistical relational A.I. Department of Computer Science and Engineering, Universiy of Washington. http;//alchemy.cs.washington.edu (2005)
9. T. N. Huynh, R. J. Mooney. Discriminative structure and parameter learning for markov logic network structure. Proc. of 25th Intl. Conf. on Machine Learning (ICML-2008) (2008).

Top-down phylogenetic tree reconstruction: a decision tree approach

Eduardo Costa[1], Celine Vens[1], and Hendrik Blockeel[1,2]

[1] Dept. of Computer Science, Katholieke Universiteit Leuven
[2] Leiden Institute of Advanced Computer Science, Universiteit Leiden
{Eduardo.Costa,Celine.Vens,Hendrik.Blockeel}@cs.kuleuven.be

A phylogenetic tree is a tree that graphically illustrates the evolutionary relationships among various species. These relationships can be inferred by analysing molecular data, such as nucleotide and amino-acid sequences. Distance matrix methods, such as Neighbor Joining (NJ) [1], are among the most popular phylogenetic tree methods. They first compute a dissimilarity measure between each pair of sequences, and then use the resulting matrix to infer a phylogenetic tree.

We propose a novel distance matrix method for reconstruction of phylogenetic trees based on a conceptual clustering method that extends the well-known decision tree learning approach [2]. Basically, our method starts from a single cluster and repeatedly divides it into subclusters until all sequences form a different cluster. Normally, there are 2^N ways to split a set of N sequences into two subsets. But if we assume that the split can be described by referring to a particular polymorphic location, then the number of splits is linear in the length of the sequences, and constant in the size of the set, making such a divisive method computationally feasible. A similar observation was made by [3], who were the first to propose a top-down clustering method for phylogenetic tree construction.

To partition a cluster into two subclusters, our method uses a criterion that is close to the optimisation criterion that NJ uses, namely, constructing a phylogenetic tree with minimal total branch length. Our algorithm is implemented as a variant of the Clus decision tree learner[3]; we call the resulting system Clus-φ.

We have tested Clus-φ on a number of synthetic datasets. The results were compared to those from NJ in terms of similarity with the true tree and computational cost. The results showed that our method scales much better in terms of the number of sequences, and remains feasible for sets of thousands of sequences, where other methods fail. Its performance is close to that of the NJ method.

Keywords: Phylogenetic trees, Bioinformatics, Decision Trees, Conceptual clustering, Top-Down phylogenetic tree

References

1. Saitou, N., Nei., M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4(4) (1987) 406-425
2. Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. In: Proc. of the 15th International Conference on Machine Learning. (1998) 55-63

[3] http://www.cs.kuleuven.be/~dtai/clus.

3. Arslan, A.N., Bizargity, P.: Phylogeny by top down clustering using a given multiple alignment. In: Proceedings of the 7th IEEE Symposium on Bioinformatics and Biotechnology (BIBE 2007), Vol. II. (2007) 809814

Using biological data to benchmark microarray analysis methods

Bertrand De Meulder, Benoît De Hertogh, Fabrice Berger, Anthoula Gaigneaux, Michael Pierre, Eric Bareke,and Eric Depiereux

Laboratoire de bioinformatique et biostatistique, Unité de Recherche en Biologie Moléculaire

Background: One key issue when analyzing microarray data is to choose (pre)treatments methods among the plethora developed. Several benchmarking methods aim to help to combine the best-suited among the millions combinations possible. These benchmarks dramatically suffer from artificial variance, far from biological relevance as they use either spike-in data or simulations. We propose a new benchmarking method avoiding those biases, by using actual biological data. We aim to take into account the actual biological variability of the data, therefore allowing finer analysis of the behavior of the statistical methods compared.

Method: We selected 34 datasets from the GEO database on the HG-U133a platform, with at least 15 replicates in each condition. Each of those datasets was pretreated with the R package GCRMA and then merged into one giant matrix (1.292.414 rows - probesets and 2*15 columns - replicates) called the DB matrix. A metric we call D/S has been computed for each row (equation 1). Then, using statistical formulae to determine a given sensitivity and positive predicting power, we computed a D/S threshold for several combinations of the number of replicates, Positive Predictive Value, and sensitivity (equation 2).

$$\frac{D}{S} = \frac{|\mu_1 - \mu_2|}{\text{mean}(S_1, S_2)} \tag{1}$$

$$\frac{D}{S_{th}} = \frac{\sqrt{2} \cdot |Z_{1-\alpha} + Z_{1-\beta}|}{\sqrt{n}} \tag{2}$$

Using this set of D/S thresholds, we resampled randomly subsets matrices: 200 rows above the threshold are the true positives (TP) and 19.800 rows below the threshold are the true negatives (TN). Volcano plots, MAplot and plots of variance versus average are similar to experimental microarray data. The subsets matrices, for which the truth is defined, were then analyzed using the R package PEGASE. We computed the p-values for this set of methods: Student t test and Welch t test (classic), Regularized t test, SAM and LIMMA (most cited), Shrinkage t test (recent) and finally Window [Welch] t test developed in our lab.

Results: The comparison of the methods shows a clear superiority of the methods using shrinkage of the variance estimation (Regularized t test, Shrinkage t test and Window t tests). These results confirm other benchmarks, but notable differences are observed at low number of replicates. The advantage of our method lies in the fact that virtually all the parameters of the analysis can be fine-tuned, allowing searcher to assess what methods are really suited for their particular cases.

Keywords: Microarray, Benchmarking, Biological variance

Structural Modeling of Transcriptomics Data Using Creative Knowledge Discovery

Kristina Gruden[1], Petra Kralj Novak[2], Igor Mozetič[2], Vid Podpečan[2], Matjaž Hren[1], Helena Motaln[1], Marko Petek[1], and Nada Lavrač[2]

[1] National Institute of Biology, Večna pot 111, Ljubljana, Slovenia
kristina.gruden@nib.si
[2] Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia

The overall aim of systems biology is to bring a novel perspective into understanding of complex interactions in biological systems. We present a top down approach for modeling of transcriptomics data through information fusion and creative knowledge discovery. By using onotology information as background knowledge for semantic subgroup discovery, rules are constructed that allow recognition of gene groups that are differentially expressed in different types of tissues. This information is further linked with the Biomine engine to visualize gene groups and uncover potential unexpected characteristics of the observed system. In Biomine, data from several publicly available databases were merged into a large graph and a method for link discovery between entities in queries was developed. Obtained models can thus serve as generators of research hypothesis that can be further on experimentally validated. Results of two case studies are presented to illustrate the applicability of the approach.

Keywords: transcriptomics, microarrays, creative knowledge discovery

Phenotype Prediction from Genotype Data

Giorgio Guzzetta, Giuseppe Jurman, Cesare Furlanello

Fondazione Bruno Kessler,
via Sommarive 18, 38123 Trento, Italy

Abstract. The prediction of quantitative phenotypes from genome-wide SNP data has only started to be explored as a tool for functional genomics. Lee et al. (2008) derived a Markov-chain model of complex traits in heterogeneous stock mice (http://gscan.well.ox.ac.uk/) from whole genotypic information. In this study we present the first example of a predictive pipeline based on Support Vector Regression and challenge Lee's results on the same GSCAN data. Comparable or better accuracies are found on two quantitative phenotypes, 12112 SNPs and about 1500 samples. A comparison with candidate loci previously identified by standard association studies also shows good agreement. Further regularization methods have been implemented, based on the naïve elastic net (Zhou and Hastie 2005, De Mol et al 2009), with prediction accuracy comparable to both SVR and the Markov-chain methods.

Keywords: GWAS, Support Vector Regression, functional genomics

In this work we propose a machine learning regression approach for genome-to-phenotype prediction to support the use of quantitative phenotypes as target variables in functional genomics.

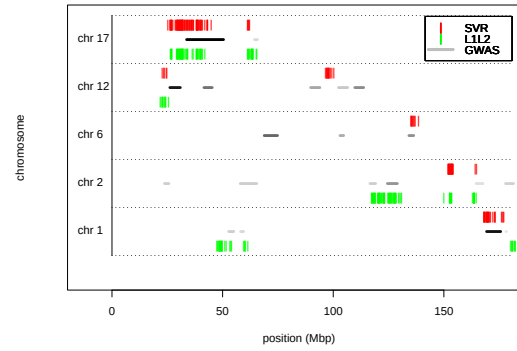
Quantitative phenotypes emerge everywhere in systems biology and biomedicine. They are of special interest in complex common diseases in which high individual variability makes difficult or impossible to separate cases into distinct categories. Fitting quantitative phenotypes from genome-wide data was only recently considered [1]. It is indicated as a promising tool when a pathophysiologic state depends on multiple genetic alterations, and especially in studying gene-environment interaction, where the effect of risk factors can be modified by environmental exposure on a multiplicity of genes [2].

While most of the machine learning studies on molecular data are focusing on classification, here we explore the use of different regularization methods in the prediction of quantitative phenotypes. We have implemented a complete pipeline, also available for high performance computing facilities, that can be applied on high-density genotype data. We present examples of prediction of quantitative phenotypes on a genome-wide dataset of 12K SNPs, comparing to Reversible Jump Monte Carlo Markov Chain (MCMC) [1]. We considered first a standard Support Vector Regression (SVR) algorithm and then the L1L2 Regression [3]. For the former, we used ε-SVR in the `LIBSVM` implementation (`http://www.csie.ntu.edu.tw/~cjlin/libsvm`),with features ranked according to their regression weights. The L1L2 regression [3] is a regularization method

that finds the optimal weight vector of a linear regression while maintaining a high sparsity of the solution. It is an evolution of the elastic net model [4] that includes Regularized Least Squares (RLS) regression to recursively optimize the feature weights. We have implemented a computationally efficient Python version of the L1L2, based on the `numpy` package. The L1L2 regression is used as machine learning core within an experimental pipeline implemented by functions of the `mlpy` package (`https://mlpy.fbk.eu`). A modular version of the software, implemented for use with parallel computing facilities, managed up to 550k features and a few thousands of samples.

Prediction was tested on the same data used in [1] and described in [5]. They include familiar, genotype and phenotype information from a population of heterogeneous stock mice. The phenotypes to predict from 12112 SNPs were the percentage of CD8+ cells and the mean cell haemoglobin (MCH). A Data Analysis Protocol replicating [1] (15 × 50%-training/test splits) and a textbook 15-CV were used for SVR and L1L2 respectively, with the squared correlation coefficient as error function. Results by SVR and L1L2 *vs* reference are listed in table, with similar or improved performance. Further, SNPs selected by SVR and L1L2 are consistent with GWAS [5]: see figure for CD8+.

Method	CD8+	MCH
SVR	0.306	0.147
L1L2	0.281	0.114
MCMC	0.314	0.111



References

1. Lee, S.H., van der Werf, J.H.J., Hayes, B.J., Goddard, M.E., Visscher, P.M.: Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. PLoS Genet 4(10) (2008)
2. Manolio, T.A.: Cohort Studies and Genetics of Complex Disease. Nat Genet 41(1) (2009) 5-6
3. De Mol, C., Mosci, S., Traskine, M., Verri, A.: A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data. J Comp Bio 16(5) (2009) 677-690
4. Zhou, H., Hastie, T.: Regularization and Variable Selection via the Elastic Net. J R Stat Soc B 67(2) (2005) 301-320
5. Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., et al: Genome- Wide Genetic Association of Complex Traits in Heterogeneous Stock Mice. Nat Genet 38 (2006) 879-887

Biomarker Selection by Transfer Learning with Linear Regularized Models

Thibault Helleputte, and Pierre Dupont

University of Louvain, Computing Science and Engineering Dept. & Machine Learning Group, Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium
{Thibault.Helleputte, Pierre.Dupontg}@uclouvain.be

This poster presents a novel feature selection method for classification of high dimensional data, such as those produced by microarrays. Classification of such data is challenging, as it typically relies on a few tens of samples but several thousand dimensions (genes). The number of microarray chips needed to obtain robust models is generally orders of magnitude higher than most datasets offer. The number of available datasets is however continuously rising, for example in databases like the NCBI's Gene Expression Omnibus (GEO). Building a large microarray dataset consisting of the simple juxtaposition of independent smaller datasets is difficult or irrelevant due to differences either in terms of biological topics, technical constraints or experimental protocols.

Biomarker selection specifically refers to the identification of a small set of genes, a signature, related to a pathology or an observed treatment outcome. The lack of robustness of biomarker selection has been outlined. In the context of biomarker selection from microarray data, a high stability means that different subsets of patients lead to very similar signatures and is a desirable property. The biological process explaining the outcome is indeed assumed to be mostly common among different patients.

Our feature selection technique includes a partial supervision (PS) to smoothly favor the selection of some dimensions (genes) on a new target dataset to be classified. The dimensions to be favored are previously selected with a simple univariate technique, like a t-test, from similar source datasets, for example from GEO, hence performing inductive transfer learning at the feature level. We rely here on our recently proposed PS-l2-AROM method, a feature selection approach embedded in a regularized linear model. This algorithm reduces to linear SVM learning with iterative rescaling of the input features. The scaling factors depend here on the selected dimensions on the source domains. The proposed optimization procedure smoothly favors the pre-selected features but the finally selected dimensions may depart from those to optimize the classification objective under rescaled margin constraints.

Practical experiments on several microarray datasets illustrate that the proposed approach not only increases classification performances, as usual with sound transfer learning scheme, but also the stability of the selected dimensions with respect to sampling variation. It is also shown that multiple transfer from various source datasets can bring further improvements.

Keywords: Feature Selection, Linear Models, Biomarkers, Regularization, Transfer Learning

Combining Semantic Relations from the Literature and DNA Microarray Data for Novel Hypotheses Generation

Dimitar Hristovski[1], Andrej Kastrin[2], Borut Peterlin[2], and Thomas C. Rindflesch[3]

[1] Institute of Biomedical Informatics, Faculty of Medicine, Ljubljana, Slovenia
dimitar.hristovski@mf.uni-lj.si
[2] Institute of Medical Genetics, University Medical Centre, Ljubljana, Slovenia
{andrej.kastrin, borut.peterlin}@guest.arnes.si
[3] National Library of Medicine, NIH, Bethesda, MD, USA
tcr@nlm.nih.gov

Although microarray experiments have great potential to support progress in biomedical research, results are not easy to interpret. Information about the functions and relations of relevant genes needs to be extracted from the vast biomedical literature. A potential solution is to use computerized text analysis methods.

Automatic text mining, commonly based on term co-occurrence, has been used to identify information valuable for interpreting microarray results. Here we propose the use of semantic relations (or predications) automatically extracted from the biomedical literature as a way of extending these techniques. Semantic predications convert textual content into "executable knowledge" amenable to further computation supporting research on genes and relevant diseases. In addition, we suggest that the combination of microarray data and semantic predications can profitably be exploited in the literature based discovery (LBD) paradigm to further enhance the scientific process.

We describe a method and an application that integrates semantic relations with microarray results and show its benefits in supporting enhanced access to the relevant literature for interpretation of results and novel hypotheses generation.

Keywords: micro-array data analysis, literature-based discovery

Two-Way Analysis of High-Dimensional Metabolomic Datasets

Ilkka Huopaniemi[1], Tommi Suvitaival[1], Janne Nikkilä[1,2], Matej Orešič[3], and Samuel Kaski[1]

[1] Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland
{ilkka.huopaniemi,tommi.suvitaival,janne.nikkila,samuel.kaski}@tkk.fi
[2] Department of Basic Veterinary Sciences, Division of Microbiology and Epidemiology, Faculty of Veterinary Medicine, University of Helsinki, P.O.Box 66, University of Helsinki FIN-00014 Finland
[3] VTT Technical Research Centre of Finland, P.O. Box 1000, FIN-02044 VTT, Espoo, Finland
matej.oresic@vtt.fi
http://www.cis.hut.fi/projects/mi/

We present a Bayesian machine learning method for multivariate two-way ANOVA-type analysis of high-dimensional, small sample- size metabolomic datasets. In metabolomics and other high-throughput bioinformatics studies, the data analysis task is typically differential analysis between diseased and healthy samples. This task is often further complicated by additional covariates, such as gender, treatment groups, or measurement times, requiring a multi-way analysis. The main complication is the combination of high dimensionality and low sample size, which renders classical multivariate techniques useless. We introduce a hierarchical model which does dimensionality reduction by assuming that the metabolites come in similarly-behaving, correlated groups. The key assumption is that metabolomics data has intrinsic correlations due to the existence of biochemical networks. The ANOVA-type decomposition is done on the set of reduced-dimensional latent variables, representing the correlated groups of variables. The method thus finds common up/down-regulations of clusters of metabolites, corresponding to subparts of metabolic pathways. The advantage of using Bayesian machine learning methods here is that they are not prone to overfitting and they inherently provide confidence intervals for the results. This aspect is particularly important when the number of samples is low. We apply the methods to study lipidomic profiles of a recent large-cohort human diabetes study. The study contains time-series of healthy control patients and patients that later progressed into type 1 diabetes, both including males and females. The method finds statistically significant main and interaction effects for relevant metabolite groups.

Keywords: ANOVA, factor analysis, hierarchical model, metabolomics, multi-way analysis, small sample-size

The Open and Closed-World Assumptions in Representing Systems Biology Knowledge

Agnieszka Lawrynowicz[1], Ross D. King[2]

[1] Institute of Computing Science, Poznan University of Technology, 60-965, Poznan, Poland
agnieszka.lawrynowicz@cs.put.poznan.pl
[2] Department of Computer Science, Aberystwyth University, Wales, UK
rdk@aber.ac.uk

We investigate knowledge representation in systems biology. We find that there is often an inconsistency between the semantics implied by the representation used to store the knowledge, and the way the knowledge is used in machine learning. We focus on the open and closed-world assumptions and their use in two important problem areas: predicting gene functional classes, and learning in metabolic networks.

The problem of learning to predict the gene functional classes is technically interesting in machine learning because of the hierarchical structure of the classes imposes dependencies on the predictions, and because a gene may have more than one function [1]. The bioinformatic data for this problem is typically represented using OWL, which makes sense given its hierarchical structure. However, when the data is pre-processed for machine learning, the open-world assumption of OWL is replaced by a closed-world assumption: if an example is not labeled as being in a class then it is a negative example. This is generally true, and it enables induction to be much more efficient, but it also introduces errors into the knowledge base.

A similar situation occurs with the interesting machine learning problems associated with metabolic pathways. Again OWL is typically used in the bioinformatics databases. However, when learning metabolic models it is more biologically natural to assume a closed-world assumption: for example if you remove the only gene encoding an enzyme that makes metabolite X then X will not be present [2,3]. As in the functional genomics example, the closed-world assumption introduces errors but has great utility for learning.

We conclude that it is important to make explicit any assumptions made when storing or reasoning with systems biology knowledge.

Keywords: predicting gene function, metabolic networks

References

1. Clare, A.J. & King, R.D. (2002) Machine learning of functional class from phenotype data. Bioinformatics 18, 160-166.
2. King, R.D., et al. (2009) The Automation of Science. Science. 324, 85-89
3. Alan Ruttenberg, A., Rees, J., Zucker, J. (2006) What BioPAX Communicates and how to extend OWL to help it. Proceedings of the OWLED 2006 Workshop

Learning gene networks with sparse inducing estimators

Fabian Ojeda, Marco Signoretto, and Johan Suykens

Department Electroteckniek, Katholieke Universiteit Leuven, Belgium
{fabian.ojeda,marco.signoretto,johan.suykens}@esat.kuleuven.be

Graphical models have drawn interest in many fields due to visualization and interpretation capabilities for complex networks. In molecular biology for instance, large-scale sparse graphs arise naturally in gene regulatory networks (GRN). If nodes in the graph are considered as random variables and edges represent interactions, the theory of probability enables the analysis of the dependence structure in the network. Gaussian graphical models [4] are an attractive choice, in virtue of simple interpretation and well established statistical theory.

The present study explores and compares different methodologies to infer the dependence structure for networks of p genes, based upon n ¡¡ p observations (expression levels). In particular, we look into the class of regression-based approaches [6, 3, 7]. In the simplest case, the inference results in the independent estimation of a regression model for each variable (gene) [6]. Sparse-inducing estimators (such as LASSO), are used to enforce that few coefficients in the estimated models are actually non-zero. In turn, as these coefficients are associated to edges in the Gaussian graphical model, this amounts at imposing a sparsity prior over the set of edges. However, the estimated structure might be self-contradictory as symmetry of the interactions is not guaranteed. To overcome this, recent work suggests to estimate a joint regression model [7]. Remarkably, joint estimation is somewhat closer to another class of structure-learning methodologies based on convex optimization procedures [2, 1, 8].

Altogether these methods have shown to be promising especially for high-dimensional problems whereas other approaches are not applicable. However, just as any other penalized estimators, the mentioned procedures face the same dilemma: the selection of regularization parameter(s). We thus investigate on practical cases the different existing criteria to accomplish model selection. Even though the chosen parameters are meaningful for the specific task, the estimated structure is generally highly unstable as n ¡¡ p. Slight changes in the training set might lead to significant changes in the estimated network. Therefore we look into the problem of stabilizing the estimation process. A plausible way to deal with this for GRN, is that of introducing a prior derived from additional gene-gene information available in public repositories. Alternatively, networks obtained from independent datasets can be combined to find a more consistent structure. Simulations on both synthetic data and gene expression data for Saccharomyces cerevisiae are reported. The regulatory program for yeast organisms is well documented and therefore, the outcome of the different algorithms can be systematically validated.

Keywords: Graphical models, gene networks, LASSO, convex-optimization

References

1. O. Banerjee and G. Natsoulis. Convex optimization techniques for fitting sparse gaussian graphical models. In Proceedings of the 23rd international conference on Machine learning, pages 89-96. ACM New York, NY, USA, 2006.
2. J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 2007.
3. T. IMOTO and R. MIYANO. Weighted lasso in graphical gaussian modeling for large gene network estimation based on microarray data. Genome Informatics 2007, page 142, 2007.
4. S. Lauritzen. Graphical models. Oxford University Press, USA, 1996.
5. C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics, 24(9):1175, 2008.
6. N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the Lasso. Annals of statistics, 34(3):1436, 2006.
7. J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial Correlation Estimation by Joint Sparse Regression Models. Journal of the American Statistical Association,
   104(486):735-746, 2009.
8. M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. Biometrika, 94(1):19-36, 2007.

Taking Advantage of the Amount of Archived Affymetrix GeneChips to Identify Genes Involved in Metastasis and Regulated by Hypoxia

Michael Pierre[1], Anthoula Gaigneaux[1], Bertrand DeMeulder[1], Fabrice Berger[1], Benoît DeHertogh[1], Eric Bareke[1], Carine Michiels[2], and Eric Depiereux[1]

[1] Molecular Biology Research Unit (URBM), University of Namur - FUNDP, Belgium
[2] Cell Biology Research Unit (URBC), University of Namur - FUNDP, Belgium
michael.pierre@fundp.ac.be

Background: Growing tumors are characterized by the presence of hypoxic areas at their center due to the lack of oxygen. It is now accepted that hypoxia selects cancer cells able to metastasize. However, the genes involved in this process remain largely unidentified. Lots of experiments using DNA microarrays have been published. However, lots of these datasets have not been fully exploited. This work proposes a methodology for the meta-analysis of several Affymetrix datasets.Applied to metastasis and/or hypoxia datasets, this methodology identified genes already known to be involved in metastasis and/or hypoxia as well as new ones.

Methods: 22 Affymetrix datasets about metastasis and/or hypoxia were downloaded from GEO and ArrayExpress. AffyProbeMiner's CDFs were applied to the CEL files. GCRMA was used for the pre-processing steps. The datasets were processed with the Window Welch t test. Based on the results of these individual analyses, two approaches were followed to select genes of interest. Another approach consisted to group several datasets into meta-datasets on which regular analyses were run.

Results: A total of 183 genes were common to two approaches or more. Out of these 183 genes, 99, such as JUNB, FOS and TP63, are already known to be involved in cancer. Moreover, 39 genes of those, such as SERPINE1 and MMP7, have been described in the literature to regulate metastasis. Twenty-one genes among which VEGFA and ID2 have also been described to be involved in the response to hypoxia. Lastly, DAVID classified those 183 genes in 24 pathways, among which 8 are directly related to cancer while 5 others are related to proliferation and cell motility. A negative control composed of 183 random genes failed to provide such results. Interestingly, 6 pathways retrieved by DAVID with the 183 genes of interest concern pathogen recognition and phagocytosis.

Keywords: Microarray,Meta-analysis, Metastasis, Hypoxia, Cancer

Metabolic syndrome assessment using Fuzzy Artmap neural network and 1H NMR spectroscopy

Bogdan Pogorelc[1,2], Jesus Brezmes[3], Matjaž Gams[1,2]

[1] Jožef Stefan Institute, Department of Intelligent Systems, Jamova c. 39, 1000 Ljubljana
[2] Špica International d.o.o., Pot k sejmišču 33, 1231 Ljubljana, Slovenia
[3] MINOS, Department Enginyeria Electronica, Universitat Rovira i Virgili, Avda. Paisos Catalans 26, 43007 Tarragona, Spain
bogdan.pogorelc@ijs.si

The metabolic syndrome is a combination of medical disorders that increase one's risk for diabetes (type 2) or cardiovascular disease, such as atherosclerosis. Due to unhealthy way of living it affects a large number of people in the developed countries. In some studies, the prevalence in the USA is calculated as being up to 25% of the population.

The presence of a metabolic syndrome relates to particular changes in the lipoprotein subclass profile, but the current clinical risk assessment methodology cannot take this properly into account. The use of Proton Nuclear Magnetic Resonance (1H NMR) spectroscopy seems to have significant potential in clinical use since the technique enables fast measurement of the lipoprotein profile directly from a serum sample. Although many computational methods have been developed to exploit these data for assessment of metabolic syndrome, a probabilistic approach giving confidence value for each serum sample measured and classified is rarely used.

The main purpose of this study was to implement the Fuzzy Artmap neural network in ANSI C to classify 1H NMR serum spectra according to the health state of individuals and to obtain the confidence value for each classified serum sample.

The approach was evaluated on extensive simulated dataset of spectra that represents five categories of health state of individuals depending on their lipoprotein subclass. Categories were: (i) healthy controls, (ii) metabolic pathway 1 (near healthy), (iii) metabolic pathway 2, (iv) metabolic pathway 3 (near metabolic syndrome) and (v) metabolic syndrome. The dataset was built based on experimental lipoprotein subclass information and comprised 2500 spectra; half of them were used for training and the other half for independent evaluation. Voting strategy was applied to present training and evaluation data 50 times in random order, giving a statistically representative output of classification with a confidence value. The results showed that the proposed approach is capable of correctly classifying 94% of the spectra. Moreover, from the confusion matrix it is evident that confusion only occurred between adjacent categories (e.g., 7.3% of healthy individuals were erroneously classified as belonging to the metabolic pathway 1 category).

In line with other recent studies our method also confirmed that the use of 1H NMR spectroscopy for metabolic syndrome and atherosclerosis risk assessment is feasible, especially since our simulated spectra included realistic statistical population variation in the lipoprotein subclass signals.

Our results show that the proposed approach can be used as an efficient tool for automatic assessment of metabolic syndrome from 1H NMR serum spectra. The Fuzzy Artmap neural network has shown its ability to classify samples with a confidence value; this is an important issue regarding the potential use of 1H NMR spectroscopy in disease risk assessment.

Keywords: Fuzzy Artmap,neural network, 1H NMR, metabolic syndrome spectroscopy, disease risk assessment

Modeling phagocytosis - PHAGOSYS project outline

Barbara Szomolay

Imperial College London

Current mathematical models of phagocytosis are mostly focusing on Rho GTPases [3], [4] and on macrophage efficacy [5], [6]. A recent modeling paper by Zerial et al. [7], suggested a 'cut-out-switch' behavior of Rab5-Rab7 conversion. However, none of these models address the question how pathogenic bacteria alters phagosome maturation and what pathways are inhibited.

Antibiotics typically target the pathogen, rather than host-specific pathways. In [1], kinase inhibitors have been developed that prevent intracellular growth of S. Typhymurium and M. tuberculosis. Akt1 has been identfied as a master regulator by controlling at least 2 essential host pathways - PAK4-RAC1/RHOA and AS160-Rab14, which can be manipulated by various pathogens.

A recent study [2] by the same authors links Akt1 with Irgm1, a member of the IFN-gamma regulated GTPase. This link is somewhat controversial as activation of Irgm1 supports, but phosphorylated Akt1 inhibits the elimination of pathogens through Rab14 activation. Like the Rab5-Rab7 switch, this could be another modeling problem of a competition between Irgm1 and Rab14 GTPases. In light of this example, we aim to construct mathematical models that will address possible crosstalks based on the literature between GTPases and manipulate them by mimicking the pathogen. One of our goals is to determine which of the different reaction schemes proposed are consistent with the dynamics of the solutions.

Keywords:

References

1. Kuijl et al., Nature 450, 2007
2. Kuijl & Neefjes, Nature Immunology 10, 2009
3. Edelstein-Keshet et al., Bull. Math. Biol. 68,69 2006, 2007
4. Goryachev et al., PLOS 2, 2006
5. Macura et al., Infection & Immunity 75, 2007
6. Jordao et al., Cell. Microbiol. 10, 2008
7. Zerial et al., Mol. Sys. Biol. 4, 2008

Inductive Process-Based Modeling of Endocytosis from Time-Series Data

Ljupčo Todorovski[1] and Sašo Džeroski[2]

[1] University of Ljubljana, Gosarjeva 5, SI-1000 Ljubljana, Slovenia
ljupco.todorovski@fu.uni-lj.si
[2] Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
saso.dzeroski@ijs.si

Inductive process-based modeling [1] is an approach to automated learning of models of dynamic systems from time-series data. The approach takes as input modeling knowledge from the domain at hand and time course measurements of the observed system. The knowledge is represented in terms of generic entities and processes in the domain; in the domain of modeling the dynamics of biological (metabolic) networks, entities correspond to metabolites, while processes denote biochemical reactions among metabolites that influence their concentrations. The learned process-based model identifies the set of specific processes among specific entities that govern the behavior of the observed system. When simulated, the learned model should closely match the time course data provided as input. Inductive process-based modeling has been successfully applied in the domain of system biology to learn models of the dynamics of biological networks; for an overview of these applications, see [3].

In this paper, we address the task of modeling endocytosis, more specifically the maturation of endosomes, a type of membrane-bound intracellular compartments. We use process-based modeling to replicate the results reported in [2], where the focus is on explaining the GTPase mechanism that switches between transporting and degrading cargo in endosomes. The switch is achieved through the replacement of Rab5 domain proteins in early endosomes with Rab7 domain proteins in mature ones. Two modeling alternatives for the mutual exclusiveness of Rab5 and Rab7 are considered: toggle vs. cut-out switch. These are compared in the context of 23 time series of Rab5 concentrations, measured by tracking early endosomes in three different experiments. The comparison shows that the model based on the cut-out switch is better supported by empirical evidence [2].

To apply the process-based modeling approach to the task of modeling the dynamics of Rab5-to-Rab7 conversion in endocytosis, we first encoded a library of domain-specific knowledge. The library includes definitions of generic entities (proteins and concentrations thereof as entity properties) and generic processes. The processes involved in the modeled GTPase mechanism include activating and inhibitory interactions, processes of exchange and hydrolysis that shuffle between inactive GDP-bound and active GTP-bound conformations of Rab5 and Rab7 proteins, and definitions of hyperbolic (Michaelis-Menten) and sigmoidal (Hill) kinetic laws for modeling individual interactions.

The library of generic entities and generic processes, together with the time-series data sets of the dynamic change of Rab5 concentration through time, was taken as input to the HIPM tool [4] for inductive process-based modeling. HIPM successfully reconstructed in an automated fashion the manually constructed Rab5-to- Rab7 conversion models [2], thus clearly demonstrating the utility of our approach.

Keywords: inductive process modeling, endocytosis

Acknowledgments. We acknowledge the support of the Phagosys (Systems Biology of Phagosome Formation and Maturation - Modulation by Intracellular Pathogens) project, funded by the European Commission under contract HEALTH-2007- 223451.

References

1. Bridewell, W., Langley, P., Todorovski, L., Džeroski, S.: Inductive process modeling. Machine Learning 71, 1-32 (2008)
2. Del Conte-Zerial, P., Brusch, L., Rink, J.C., Collinet, C., Kalaidzidis, Y., Zerial, M., Deutsch, A.: Membrane identity and GTPase cascades regulated by toggle and cut-out switches. Molecular Systems Biology 4, doi:10.1038/msb.2008.45 (2008)
3. Džeroski S., Todorovski L.: Equation discovery for systems biology: finding the structure and dynamics of biological networks from time course data. Current Opinion in Biotechnology 19(4), 360-368 (2008)
4. Todorovski, L., Bridewell, W., Shiran, O., Langley, P.: Inducing hierarchical process models in dynamic domains. Proceedings of the Twentieth National Conference on Artificial Intelligence, 892-897 (2005)

Analyzing time series gene expression data with predictive clustering rules

Bernard Ženko[1], Jan Struyf[2], and Sašo Džeroski[1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute Jamova cesta 39, SI-1000 Ljubljana, Slovenia
{bernard.zenko, saso.dzeroski}@ijs.si
[2] Department of Computer Science, Katholieke Universiteit Leuven Celestijnenlaan 200A, B-3001 Leuven, Belgium
jan.struyf@cs.kuleuven.be

Under specific environmental conditions, co-regulated genes and/or genes with similar functions tend to have similar temporal expression profiles. Identifying groups of genes with similar temporal profiles can therefore bring new insight into understanding of gene regulation and function. The most common way of discovering such groups of genes is with short time series clustering techniques. Once we have the clusters, we can also try to describe them in terms of some common characteristics of the comprising genes, e.g., (Ernst et al., 2005). An alternative way are the so-called constrained clustering techniques; here only clusters with valid descriptions are considered, and as a result, we obtain clusters and their descriptions in one single step.

We present a novel constrained clustering method for short time series, which uses the approach of predictive clustering. Predictive clustering (Blockeel et al., 1998) combines clustering and predictive modeling; it partitions the instances in a set of clusters like the regular clustering does, however, it also constructs predictive model(s) that describes each of the clusters. So far, predictive models can take the form of decision trees (Blockeel et al., 1998) or rules (Ženko et al., 2005). Predictive clustering trees, together with a qualitative time series distance measure (Todorovski et al., 2002), have already been used for clustering of short time series (Džeroski et al., 2007). Here we present predictive clustering rules for short time series, which use the same qualitative distance measure, but describe clusters with decision rules instead of trees.

The advantage of rules over trees is that each rule describing a cluster can be interpreted independently of other rules (clusters), while a tree describes all the clusters simultaneously. In addition, within rules we can easily introduce an additional constraint that rule conditions only comprise tests on the presence of gene descriptors and not on their absence. Trees by their nature have to include both types of tests (a set of instances is split into a cluster where the gene descriptor is present, and another set where the descriptor is absent), even if tests on absence are not biologically meaningful.

We demonstrate the benefits of our method on a publicly available collection of data sets (Gasch et al., 2000), which records the changes over time in the expression levels of yeast genes in response to a change in several environmental conditions. As the gene descriptors we use the Gene Ontology terms (Ashburner et al., 2000). The results show that rules give rise to clusters of genes with similar statistical properties (e.g., intra cluster variance and size) as trees, however, the descriptions of the clusters are easier to interpret since they only include the presences of gene descriptors.

Keywords: time series, predictive clustering, rule learning

References

1. M. Ashburner et al. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, 25(1):25–29, 2000.
2. H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of clustering trees. In *15th Int'l Conf. on Machine Learning*, pages 55–63, 1998.
3. J. Ernst, Nau G.J., and Bar-Joseph Z. Clustering short time series gene expression data. *Bioinformatics*, 21(Suppl. 1):159–168, 2005.
4. L. Todorovski, B. Cestnik, M. Kline, N. Lavrac, and S. Džeroski. Qualitative clustering of short time-series: A case study of firms reputation data. In *ECML/PKDD-'2 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, pages 141–149, 2002.
5. S. Džeroski, V. Gjorgjioski, I. Slavkov, and J. Struyf. Analysis of time series data with predictive clustering trees. In *5th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers*, LNCS, Volume 4747, pages 63-80. Springer, 2007.
6. B. Ženko, S. Džeroski, and J. Struyf. Learning predictive clustering rules. In *4th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers*, LNCS, Volume 3933, pages 234–250. Springer, 2005.

Author Index

Abeel Thomas, 135, 136
Airola Antti, 15
Aloraini Adel, 137
Antal P., 35, 139, 141, 143
Atalay Volkan, 45

Böck Matthias, 144
Bareke Eric, 151, 168
Bender Andreas, 85
Berger Fabrice, 144, 151, 168
Birnie Richard, 137
Blockeel Hendrik, 85, 149
Brezmes Jesus, 169
Brouard Céline, 147

Cesa-Bianchi Nicolo', 25
Cetin-Atala Rengül, 45
Costa Eduardo, 149
Curk Tomaž, 65, 125
Cussens James, 137

d'Alché-Buc Florence, 147
Džeroski Sašo, 115, 172, 174
De Baets Bernard, 15
De Hertogh Benoît, 151
Debily Marie-Anne, 147
DeHertogh Benoît, 168
Demšar Janez, 65
DeMeulder Bertrand, 151, 168
Depiereux Eric, 151, 168
di Bernardo Diego, 1
Dubois Julie, 147
Dupont Pierre, 135, 156

Erjavec Aleš, 125

Falus A., 35, 139, 141, 143
Furlanello Cesare, 154

Gaigneaux Anthoula, 151, 168
Gams Matjaž, 169
Gorup Črtomir, 125
Gruden Kristina, 153
Guzzetta Giorgio, 154

Haj G., 139
Hanczar Blaise, 75
Helleputte Thibault, 135, 156
Hren Matjaž, 153
Hristovski Dimitar, 159
Hullám G., 35, 139
Huopaniemi Ilkka, 161

Isik Zerrin, 45

Jerala Roman, 3
Jurman Giuseppe, 154
Juty Nick, 5

Kaderali Lars, 144
Kaferle Petra, 125
Kalaidzidis Yannis, 7
Kaski Samuel, 161
Kastrin Andrej, 159
King Ross D., 9, 163
Kiszel P., 143
Kralj Novak Petra, 153
Kramer Stefan, 144

Lavrač Nada, 153
Lawrynowicz Agnieszka, 163
Leban Gregor, 65
Liu Wei, 55

Mattiazzi Mojca, 125
Michiels Carine, 168
Millinghoffer A., 139, 141
Motaln Helena, 153
Mozetič Igor, 153
Mramor Minca, 65

Nadeem Malik Sajjad Ahmed, 75
Nikkil Janne, 161
Niranjan Mahesan, 55

Ogishima Soichi, 144
Ojeda Fabian, 165
Orešič Matej, 161

Pahikkala Tapio, 15
Petek Marko, 153
Peterlin Borut, 159
Petrovič Uroš, 125
Pierre Michael, 151, 168

Podpečan Vid, 153
Pogorelc Bogdan, 169

Rahmani Hossein, 85
Re Matteo, 95
Rindflesch Thomas C., 159

Sárközy P., 143
Saeys Yvan, 105, 135, 136
Salakoski Tapio, 15
Semsei 'A., 143
Signoretto Marco, 165
Slavkov Ivica, 115
Stafford Noble William, 11
Struyf Jan, 174
Suvitaival Tommi, 161
Suykens Johan, 165
Szalai Cs., 35, 139, 141, 143
Szomolay Barbara, 171

Todorovski Ljupčo, 172
Toplak Marko, 65

Umek Lan, 125

Valentini Giorgio, 25, 95
Van de Peer Yves, 105, 135, 136
Van Landeghem Sofie, 105
Vens Celine, 149
Vrain Christel, 147

Waegeman Willem, 15

Ženko Bernard, 115, 174
Zoltán B., 143
Zucker Jean-Daniel, 75
Zupan Blaž, 65, 125

PASCAL2

Pattern Analysis, Statistical Modelling and Computational Learning

UNIVERSITY OF HELSINKI

Department of Computer Science, Faculty of Science

Series of Publications B, Report B-2009-1