ECML 2007 PKDD
WARSAW POLAND

THE 18TH EUROPEAN CONFERENCE ON MACHINE LEARNING
AND
THE 11TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE
OF KNOWLEDGE DISCOVERY IN DATABASES

---

# PROCEEDINGS OF THE 2ND WORKSHOP IN DATA MINING IN FUNCTIONAL GENOMICS AND PROTEOMICS

---

**DMFGP'07**

**September 17, 2007**

**Warsaw, Poland**

**Editors:**
*A. Fazel Famili*
National Research Council, Canada
*Xiaohui Liu*
Brunel University, UK
*José-María Peña*
Universidad Politécnica de Madrid, Spain

# Preface

Data Mining in Functional Genomics and Proteomics involves a close collaboration between researchers from a number of diverse areas, such as biology, medicine, genomics and proteomics to computer science, mathematics and statistics. This collaboration of disciplines has evolved because of the: (i) advances that have occurred in data production and acquisition facilities, such as the introduction of mircroarrays and high throughput genomics and proteomics, (ii) enormous amounts of data that is generated every day that cannot be analyzed using ordinary data mining tools and techniques, and (iii) strong interest from many groups (research institutes, hospitals, academia, pharmaceuticals, etc.) who want to benefit from this wealth of data. Many efforts to deal with these issues are being undertaken by researchers working in this field. The aim of this workshop was to bring together researchers working on different topics related to data mining in functional genomics and proteomics. In particular we were interested to focus on current trends and emphasize on what should be the future directions for generic and applied research in this field. The main topics addressed during the workshop are integration methodologies for functional genomics and proteomics and also issues related to structuring and disseminating all useful knowledge that increasingly becomes available in this field.

Our call for papers resulted in some very interesting papers that are the contents of these workshop proceedings. The workshop was organized as part of ECML/PKDD-2007 conference. We are grateful for the support that we received from the organizers of this conference. In particular we would like to thank Dr. Marzena Kryszkiewics, ECML/PKDD 2007 Workshops Chair for accepting our proposal, the program committee and additional reviewers (Drs. Amira Djebbari, Edwin Wang, and Youlian Pan) who helped us for the review.

Warsaw, September 2007

A. Fazel Famili (Chair)
Xiaohui Liu (Co-Chair)
José-María Peña (Co-Chair)

# Workshop Organization

## DMFGP Workshop Chairs

A. Fazel Famili (National Research Council – Canada) (chair)
Xiaohui Liu (Brunel University – UK) (co-chair)
José-María Peña (Universidad Politécnica de Madrid – Spain) (co-chair)

## ECML/PKDD Workshop Chair

Marzena Kryszkiewicz (Warsaw University of Technology)

## Workshop Program Committee

Guillaume Beslon (LIRIS, INSA–Lyon – France)
Henrik Bostrom (Royal Inst. Of Technology – Sweden)
Joaquin Dopazo (CIPF – Spain)
Ana Teresa Freitas (INESC-ID/IST – Portugal)
Samuel Kaski (Helsinki University of Technology -– Finland)
Alexander Schliep (Max Planck Institute — Germany)
Sergio Storari (Università di Ferrara – Italy)
Evgenii Vityaev (Russian Academy of Science – Russia)

# Table of Contents

Combining APRIORI and Bootstrap Techniques for Marker Analysis . . . . . 1
*Giacomo Gamberoni, Evelina Lamma, Fabrizio Riguzzi, Chiara Scapoli, Sergio Storari*

Discovering Informative Genes from Gene Expression Data: A Multi-strategy Approach . . . . . 11
*Fazel Famili, Sieu Phan, Ziying Liu, Youlian Pan, Amira Djebbari, Anne Lenferink, Maureen O'Connor*

Breast Cancer Biomarker Selection Using Multiple Offspring Sampling . . . . . 23
*Antonio LaTorre, José-María Peña, Santiago González, Oscar Cubo, A. Fazel Famili*

Classification of Microarrays with kNN: Comparison of Dimensionality Reduction Methods . . . . . 35
*Sampath Deegalla, Henrik Boström*

Knowledge Discovery in Neuroblastoma-related Biological Data . . . . . 45
*Edwin van de Koppel, Ivica Slavkov, Kathy Astrahantseff, Alexander Schramm, Johannes Schulte, Jo Vandesompele, Edwin de Jong, Sašo Džeroski, Arno Knobbe*

Partially-supervised Context-specific Independence Mixture Modeling . . . . . 57
*Benjamin Georgi, Alexander Schliep*

Using Symmetric Causal Independence Models to Predict Gene Expression from Sequence Data . . . . . 67
*Rasa Jurgelenaite, Tom Heskes, Tjeerd Dijkstra*

Identification of Cooperative Mechanisms in Transcription Regulatory Networks Using Non-supervised Learning Techniques . . . . . 79
*Ana T. Freitas, Ana P. Ramalho, Carlos A. Oliveira, Christian S. Nogueira, Miguel C. Teixeira, Isabel Sá-Correia, Arlindo L. Oliveira*

Generating Data from the Evolution of Artificial Regulatory Networks . . . . . 91
*Yolanda Sánchez-Dehesa, José-María Peña, Guillaume Beslon*

Transcription Factor Binding Site Discovery by the Probabilistic Rules . . . . . 104
*Irina Khomicheva, Alexander Demin, Evgeny Vityaev*

**Author Index** . . . . . 110

# Combining APRIORI and Bootstrap Techniques for Marker Analysis

Giacomo Gamberoni[1], Evelina Lamma[1], Fabrizio Riguzzi[1], Chiara Scapoli[2], and Sergio Storari[1]

[1] ENDIF, University of Ferrara, Italy
{giacomo.gamberoni, evelina.lamma, fabrizio.riguzzi, sergio.storari}@unife.it
[2] Department of Biology, University of Ferrara, Italy, scc@unife.it

**Abstract.** In genetic studies, complex diseases are often analyzed searching for marker patterns that play a significant role in the susceptibility to the disease. In this paper we consider a dataset regarding periodontitis, that includes the analysis of nine genetic markers for 148 individuals. We analyze these data by using a novel subgroup discovering algorithm, named APRIORI-B, that is based on APRIORI and bootstrap techniques. This algorithm can use different metrics for rule selection. Experiments conducted by using as rule metrics novelty and confirmation, confirmed some previous results published on periodontitis.

Full article in PDF

# Discovering Informative Genes from Gene Expression Data: A Multi-strategy Approach

Fazel Famili[1], Sieu Phan[1], Ziying Liu[1], Youlian Pan[1], Amira Djebbari[1], Anne Lenferink[2], and Maureen O'Connor[2]

[1] Institute for Information Technology, National Research Council, Ottawa Ontario, K1A 0R6, Canada;
{fazel.famili, sieu.phan, ziying.liu, youlian.pan, amira.djebbari}@nrc-cnrc.gc.ca

[2] Biotechnology Research Institute, National Research Council, Montreal, Quebec, H4P 2R2, Canada
{anne.lenferink, maureen.oconnor}@nrc-cnrc.gc.ca

**Abstract.** This paper discusses the issue of dealing with large volume of high throughput genomics data and applying unsupervised multi-strategy methods to identify differentially expressed genes. We introduce a novel method that consists of 8 steps. The approach is applied to a set of genomics data obtained from a cancer research study. We demonstrate the effectiveness of our method which includes some validation using biological experiments and literature search.

**Keywords:** Gene expression data analysis, Multi-strategy learning, Data mining and knowledge discovery.

Full article in PDF

# Breast Cancer Biomarker Selection Using Multiple Offspring Sampling

Antonio LaTorre[1], José-María Peña[1], Santiago González[1], Oscar Cubo[1], and A. Fazel Famili[2]

[1] Computer Architecture Department, Universidad Politécnica de Madrid
Boadilla del Monte, Madrid, 28660, Spain
{atorre, jmpena, sgonzalez, ocubo}@fi.upm.es
[2] Institute for Information Technology, National Research Council,
Ottawa Ontario, K1A 0R6, Canada;
fazel.famili@nrc-cnrc.gc.ca

**Abstract.** Biomarkers are biochemical facets that can be used to measure different aspects of a disease. In the last years, there has been much interest in biomarkers of different cancer variants for predicting future patterns of disease. However, DNA Biomarker selection is a difficult task as it involves dealing with a special type of datasets, microarrays, that consists of a large number of features with small number of samples. This paper proposes a new approach for biomarkers selection by means of an innovative parallel evolutionary algorithm that performs wrapper feature selection from thousands of genes to achieve a small set of most relevant ones. To test our method, the well known Van't Veer dataset on Breast Cancer has been considered. Preliminary results outperform those reported by Van't Veer both in accuracy and the number of genes selected.

Full article in PDF

# Classification of Microarrays with kNN: Comparison of Dimensionality Reduction Methods

Sampath Deegalla[1] and Henrik Boström[2]

[1] Dept. of Computer and Systems Sciences,
Stockholm University and Royal Institute of Technology,
Forum 100, SE-164 40 Kista, Sweden
si-sap@dsv.su.se
[2] School of Humanities and Informatics,
University of Skövde,
P.O. Box 408, SE-541 28, Skövde, Sweden
henrik.bostrom@his.se

**Abstract.** Dimensionality reduction can often improve the performance of the k-nearest neighbor classifier (kNN) for high-dimensional data sets, such as microarrays. The effect of the choice of dimensionality reduction method on the predictive performance of kNN for classifying microarray data is an open issue, and four common dimensionality reduction methods, Principal Component Analysis (PCA), Random Projection (RP), Partial Least Squares (PLS) and Information Gain(IG), are compared on eight microarray data sets. It is observed that all dimensionality reduction methods result in more accurate classifiers than what is obtained from using the raw attributes. Furthermore, it is observed that both PCA and PLS reach their best accuracies with fewer components than the other two methods, and that RP needs far more components than the others to outperform kNN on the non-reduced dataset. None of the dimensionality reduction methods can be concluded to generally outperform the others, although PLS is shown to be superior on all four binary classification tasks, but the main conclusion from the study is that the choice of dimensionality reduction method can be of major importance when classifying microarrays using kNN.

Full article in PDF

# Knowledge Discovery in Neuroblastoma-related Biological Data

Edwin van de Koppel[1], Ivica Slavkov[2], Kathy Astrahantseff[3], Alexander Schramm[3], Johannes Schulte[3], Jo Vandesompele[4], Edwin de Jong[1], Sašo Džeroski[2], and Arno Knobbe[1]

[1] Utrecht University, P.O. box 80 089, NL-3508 TB Utrecht, The Netherlands, Department of Knowledge Technologies
[2] Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia2, Center of Pediatric Oncology
[3] University Children's Hospital of Essen, Hufelandstrasse 55, 45122 Essen, Germany
[4] University of Ghent, Centre for Medical Genetics, De Pintelaan185, 9000, Gent, Belgium
Contact: a.knobbe@kiminkii.com

**Abstract.** In this paper, we provide initial Data Mining results on four sets of genetic data, collected in the context of the new European Embryonal Tumour Pipeline project. These data sets provide different views on the genetic processes involved in the genesis and development of a specific type of tumour, known as neuroblastoma. Although the project involves other types of tumours as well, with potentially similar underlying causal processes, neuroblastoma is currently the only disease for which sufficient data has been collected to analyse. We provide results on this data using systems developed at two Data Mining groups in Europe, with the aim of introducing the different Data Mining challenges involved, and outlining the approach we intend to apply throughout the project. Our descriptions focus on the analysis of individual data sets, stemming from separate analysis platforms (e.g. Affymetrix microarrays). Additionally, we provide some pointers for doing cross-platform analysis in the future.

Full article in PDF

# Partially-supervised Context-specific Independence Mixture Modeling

Benjamin Georgi and Alexander Schliep

Max Planck Institute for Molecular Genetics, Dept. of Computational Molecular Biology,
Ihnestrasse 73, 14195 Berlin, Germany

**Abstract.** Partially supervised or semi-supervised learning refers to machine learning methods which fall between clustering and classification. In the context of clustering, labels can specify *link* and *do-not-link* constraints between data points in different ways and constrain the resulting clustering solutions. This is a very natural framework for many biological applications as some labels are often available and even very few labels greatly improve clustering results.
Context-specific independence models constitute a framework for simultaneous mixture estimation and model structure determination to obtain meaningful models for high-dimensional data with many, possibly uninformative, variables. Here we present the first approach for partial learning of CSI models and demonstrate the effectiveness of modest amounts of labels for simulated data and for protein sub-family determination.

Full article in PDF

# Using Symmetric Causal Independence Models to Predict Gene Expression from Sequence Data

Rasa Jurgelenaite[1], Tom Heskes[1], and Tjeerd Dijkstra[2]

[1] Institute for Computing and Information Sciences, Radboud University Nijmegen,
PO Box 9010, 6500 GL Nijmegen, The Netherlands
{rasa, tomh}@cs.ru.nl
[2] Department of Parasitology, Leiden University Medical Center,
PO Box 9600, 2300 RC Leiden, The Netherlands
t.dijkstra@lumc.nl

**Abstract.** We present an approach for inferring transcriptional regulatory modules from genome sequence and gene expression data. Our method, which is based on symmetric causal independence models, is both able to model the logic behind transcriptional regulation and to incorporate uncertainty about the functionality of putative transcription factor binding sites. Applying our approach to the deadliest species of human malaria parasite, Plasmodium falciparum, we obtain several striking results that deserve further (biological) investigation.

**keywords**: Transcriptional regulatory networks, symmetric causal independence models, Plasmodium falciparum

Full article in PDF

# Identification of Cooperative Mechanisms in Transcription Regulatory Networks Using Non-supervised Learning Techniques

Ana T. Freitas[1], Ana P. Ramalho[1], Carlos A. Oliveira[1], Christian S. Nogueira[1], Miguel C. Teixeira[2], Isabel Sá-Correia[2], and Arlindo L. Oliveira[1]

[1] INESC-ID/IST, Lisboa, Portugal.
[2] IBB-CEBQ/IST, Lisboa, Portugal

**Abstract.** Accurate identification of biological processes and gene regulatory mechanisms remains an open problem that needs to be addressed, since it represents a key stepping stone in our path to the goal of systems biology. In this work, we propose a new methodology for the identification of putative cooperative regulatory mechanisms in transcription regulatory networks.

Our approach is based on the identification of biclusters in a binary regulation matrix, where each row corresponds to a given transcription factor, each column to a target gene and each matrix entry contains the value one if there is a documented (or potential) regulation between the transcription factor and the gene that correspond to that entry.

We show that a perfectly homogeneous bicluster in this matrix identifies a set of transcription factors that jointly regulate a group of genes. Less than perfectly homogeneous biclusters may also provide new clues about still unknown regulation processes. Our methodology was tested using data from a microarray analysis of the early response of the eukaryotic model *Saccharomyces cerevisiae* to the widely used herbicide 2,4-D dichlorophenoxyacetic acid.

Results obtained with this validation show that the approach has the ability to uncover groups of genes and transcription factors that are jointly involved in a number of relevant biological processes. More significantly, our method has the potential to identify a number of possibly interesting new hypotheses of regulatory mechanisms, that may be validated experimentally.

Full article in PDF

# Generating Data from the Evolution of Artificial Regulatory Networks

Yolanda Sánchez-Dehesa[1], José-María Peña[2], and Guillaume Beslon[1]

[1] Laboratoire d'InfoRmatique en Images et Systémes d'information
UMR 5205 CNRS / INSA de Lyon/Université Claude Bernard Lyon 1 / Université Lumiére Lyon 2 / Ecole Centrale de Lyon, France
[2] Computer Architecture Department, Universidad Politécnica de Madrid
Boadilla del Monte, Madrid, 28660, Spain

**Abstract.** Existing regulatory network models attempt to copy the "in vivo" regulatory principles by reproducing founded biological results "in silico". These models sometimes don't reflect the biological principal of protein regulation and they don't take into account the organisms evolution. An innovative approach is presented on this contribution, based on the analyze of the existing models. Biological principles of regulatory networks have been considered. In fact, this new model will provide the tools to study regulatory networks emergency and evolution and to acquire knowledge from generated time series. Studying networks in silico provide us the tools for controlling the environment and a better behavior analysis.

**keywords**: evolutionary algorithm, regulatory networks, protein transcription, network emergency, dynamical properties, artificial data generation

Full article in PDF

# Transcription Factor Binding Site Discovery by the Probabilistic Rules

Irina Khomicheva[1], Alexander Demin[2], and Evgeny Vityaev[3]

[1] Institute of Cytology and Genetics SB RAS, Lavrentyev aven., 10 630090 Novosibirsk, Russia
khomicheva@bionet.nsc.ru
[2] A.P.Yershov Institute of Information Systems SD RAS, Lavrentyev aven., 6, 630090 Novosibirsk, Russia
alexandredemin@yandex.ru
[3] Sobolev Institute of Mathematics, Koptyug aven. 4, 630090 Novosibirsk, Russia
vityaev@math.nsc.ru

**Abstract.** Control of gene expression at the level of transcription is achieved by nuclear factors that bind to regulatory elements, short DNA sequence motifs, called transcription factor binding sites. The development of reliable methods for binding site recognition is an important step in large-scale genome analysis. The Data Mining approaches adapted to bioinformatics tasks show high efficiency. Yet the specificity of the regulatory region analysis task consists in the high false-positive rates. In this paper the program system 'Discovery' was applied to tasks of binding site recognition. 'Discovery' makes a semantic probabilistic inference and finds the statistically significant probabilistic rules. The hypothesis class is defined by the expert in dialog mode. In this paper we demonstrate that 'Discovery' is consistently more accurate than the traditional weight matrices in binding site prediction task, as was established for three families of transcription factors.

**Keywords**: semantic probabilistic inference, probabilistic rule, transcription factor binding sites, prediction.

Full article in PDF

# Author Index


Astrahantseff, Kathy, 45

Beslon, Guillaume, 91
Boström, Henrik, 35

Cubo, Oscar, 23

Džeroski, Sašo, 45
de Jong, Edwin, 45
Deegalla, Sampath, 35
Demin, Alexander, 104
Dijkstra, Tjeerd, 67
Djebbari, Amira, 11

Famili, A. Fazel, 11, 23
Freitas, Ana T., 79

Gamberoni, Giacomo, 1
Georgi, Benjamin, 57
González, Santiago, 23

Heskes, Tom, 67

Jurgelenaite, Rasa, 67

Khomicheva, Irina, 104
Knobbe, Arno, 45

Lamma, Evelina, 1
LaTorre, Antonio, 23
Lenferink, Anne, 11
Liu, Ziying, 11

Nogueira, Christian S., 79

O'Connor, Maureen, 11
Oliveira, Arlindo L., 79
Oliveira, Carlos A., 79

Pan, Youlian, 11
Peña, José-María, 23, 91
Phan, Sieu, 11

Ramalho, Ana P., 79
Riguzzi, Fabrizio, 1

Scapoli, Chiara, 1
Schliep, Alexander, 57
Schramm, Alexander, 45
Schulte, Johannes, 45
Slavkov, Ivica, 45
Storari, Sergio, 1
Sá-Correia, Isabel, 79
Sánchez-Dehesa, Yolanda, 91

Teixeira, Miguel C., 79

van de Koppel, Edwin, 45
Vandesompele, Jo, 45
Vityaev, Evgeny, 104