

ECML/PKDD 2004

15th European Conference on Machine Learning (ECML)
8th European Conference on Principles and Practice
of Knowledge Discovery in Databases (PKDD)

Pisa, September 20 - 24, 2004



Proceedings of the Workshop W9 on Data Mining and Text Mining for Bioinformatics

Editor
Tobias Scheffer (Humboldt University, Berlin, Germany)

Preface

In the past years, research in molecular biology and molecular medicine has accumulated enormous amounts of data. This includes genomic sequences gathered by the Human Genome Project, gene expression data from microarray experiments, protein identification and quantification data from proteomics experiments, and SNP data from high-throughput SNP arrays. However, our understanding of the biological processes underlying these data lags far behind. There is a strong interest in employing methods of knowledge discovery and data mining to generate models of biological systems. Mining biological databases imposes challenges which knowledge discovery and data mining have to address, and which form the focus of the European Workshop on Data Mining and Text Mining for Bioinformatics.

This volume contains the papers presented at the Second European Workshop on Data Mining and Text Mining for Bioinformatics, held at the European Conference on Machine Learning and the European Conference on Principles and Practice of Knowledge Discovery in Databases, in Pisa, Italy, on September 24, 2004. Two invited and ten contributed papers were presented at the workshop; invited presentations were given by Yves Moreau and Alfonso Valencia.

I would like to thank the members of the program committee for their great help in the reviewing process. I would also like to thank the European Network of Excellence in Knowledge Discovery (KD-Net) for financial support.

Berlin, 2004.

Tobias Scheffer

Organization

Program Chair

Tobias Scheffer, Humboldt-Universität zu Berlin, Germany.

Program Committee

Sourav Bhomwick, Nanyang Technological University, Singapore.

Christian Blaschke, Centro Nacional de Biotecnología.

Vladimir Brusilovskiy, Institute for Infocomm Research, Singapore.

Carol Friedman, Columbia University.

George Forman, Hewlett Packard.

Rob Gaizauskas, University of Sheffield.

Jrg Hakenberg, Humboldt-Universität zu Berlin.

Ross King, University of Wales, Aberystwyth, and PharmaDM.

Adam Kowalczyk, Telstra & Peter MacCallum Cancer Centre.

Stefan Kramer, Technische Universität München.

Ulf Leser, Humboldt-Universität zu Berlin.

Bhavani Raskutti, Telstra.

Steffen Schulze-Kremer, Max-Planck-Institute, Berlin.

Myra Spiliopoulou, University of Magdeburg.

Alfonso Valencia, Centro Nacional de Biotecnología, Spain.

David Vogel, AI Insight.

Mohammed Zaki, Rensselaer Polytechnic Institute.

Sponsors

The workshop received support from the European Network of Excellence in Knowledge Discovery (KD-Net).

Table of Contents

Invited Papers

Integrating Text and Microarray Data: Gene Expression and Comparative Genomic Hybridization
Yves Moreau 1

Information Extraction in Molecular Biology
Alfonso Valencia 2

Contributed Papers

An Evaluation of Gene Selection Methods for Multi-Class Microarray Data Classification
Hong Chai and Carlotta Domeniconi 3

Extracting Protein Function Information from MEDLINE Using a Full-Sentence Parser
Nikolai Daraselia, Sergei Egorov, Andrey Yazhuk, Svetlana Novichkova, Anton Yuryev, and Ilya Mazo 11

A Method of Extracting Sentences Related to Protein Interaction from Literature using a Structure Database
Yoshikazu Kaneta, Md. Ahaduzzaman Munna, and Takenao Ohkawa 18

Hierarchical Text Categorization as a Tool of Associating Genes with Gene Ontology Codes
Svetlana Kiritchenko, Stan Matwin, and A. Fazel Famili 26

Duplicate Detection in Biological Data using Association Rule Mining
Judice L.Y.Koh, Mong Li Lee, Asif M. Khan, Paul T.J. Tan, and Vladimir Brusic 31

Assessment of SVM Reliability for Microarrays Data Analysis
Andrea Malossini, Enrico Blanzieri, Raymond T. Ng 38

Five Steps to Text Mining in Biomedical Literature
Brigitte Mathiak and Silke Eckstein 43

Protein Fold Recognition with K-Local Hyperplane Distance Nearest Neighbor Algorithm
Oleg Okun 47

A Finite State Automata Based Technique for Protein Classification Rules Induction
F. Psomopoulos, S. Diplaris, and P. A. Mitkas 54

Mining relations in the GENIA corpus
Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, James Dowdall, Andreas Persidis, and Ourania Konstanti 61

Integrating Text and Microarray Data: Gene Expression and Comparative Genomic Hybridization

Yves Moreau

Katholieke Universiteit Leuven ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven,
yves.moreau@esat.kuleuven.ac.be

Microarrays are now a well established method for gene expression analysis, while another type of microarrays - Comparative Genomic Hybridization (CGH) microarrays - is an emerging technology for the detection of chromosomal aberrations. First, we consider the integration of expression data directly with literature information, which is still much of an unsettled issue. Second, we consider the new questions raised by the linking of microarray CGH data with textual patient cases to discover links between genetic aberrations and clinical phenotypes.

For gene expression data, we first consider the vector-based text representation, which compiles a term-based profile for each gene based on its associated literature. Such gene profiles capture relevant functional information, which we automatically assess through a text-based score of the functional cohesion of a group of genes. Through a use of multiple, complementary vocabularies, the public text mining web service TXTGate offers several views on gene literature and it is geared at profiling groups of genes, typically arising from cluster analysis. Second, we combine gene expression profiles and text profiles directly in a clustering framework based on statistical meta-analysis. Because literature descriptions of genes and gene expression data offer overlapping but distinct information of the relationship between genes, improvement of clustering results can be expected. As many parameter settings need to be evaluated in parallel, we propose a score that gives a one-shot estimate of the enrichment of known transcription factor binding sites in the meta-clustering solution. Through this independent validation, we demonstrate the overall benefit to use expression data and literature jointly when clustering gene expression data.

CGH microarray is an emerging microarray extension of classical CGH that allows genomewide identification of submicroscopic chromosomal deletions and amplifications (through hybridization of genomic patient DNA to an array of clones tiling the human genome). By cross-correlating information from textual case reports from multiple patients to the corresponding aberrant genomic locations, it is possible to identify candidate genes for the phenotypic features linked to a group of cases. CGHGate is a database and text-mining tool to visualize data on a genomic region of interest (e.g., via a DAS Server (Distributed Annotation Server) on Ensembl). It allows retrieval of case reports resembling other patients with similar aberrations, and to biomedical literature (linking to MEDLINE abstracts). These text-mining features are based on the vector space model and IDF (Inverse Document Frequency) indices of case reports and MEDLINE abstracts, using tailored controlled vocabularies that are domain specific (i.e., gene, disease, and dysmorphology centric). The generation of a phenotypic genome map combined with text mining features will enable the identification of genes involved in developmental processes and will delineate novel clinically recognizable entities.

Information Extraction in Molecular Biology

Alfonso Valencia

Protein Design Group, National Center for Biotechnology, CNB-CSIC Cantoblanco, Madrid E-28049,
valencia@cnb.uam.es, <http://www.pdg.cnb.uam.es>

Integrating vast amounts of heterogeneous data, generated by Genomics and Proteomics approaches, and linking it with the vast amount of detailed information accumulated in scientific publications is one of the more interesting challenges of bioinformatics (see Valencia, EMBO Reports, 2002).

Current areas of development include: access and organization of the textual information. development of comprehensive repositories of annotated text in the various knowledge domains; identification of entities in text (e.g. protein and gene names, diseases, drugs, species, tissues and others), accurate description of the relations between entities (e.g. protein interactions, gene control relations and. genes disease relations), and global relations (e.g. function common to a set of genes), and representation of the extracted knowledge (e.g. graphical formats, database querying capabilities). For a review see, Blaschke et al., Brief. in Bioinform. 2002. The highly specialized nature of the research in Molecular Biology makes necessary the development of specific tools and applications what have resulted in the proliferation of tools and applications. In this situation it is very important to compare and assess competing approaches with common standards and evaluation criteria (see the description of the BioCreative competition at <http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.htm>).

In this presentation I'll review the situation of the field, the main challenges still ahead of us, and my view on the impact in these developments in development of Molecular Biology.