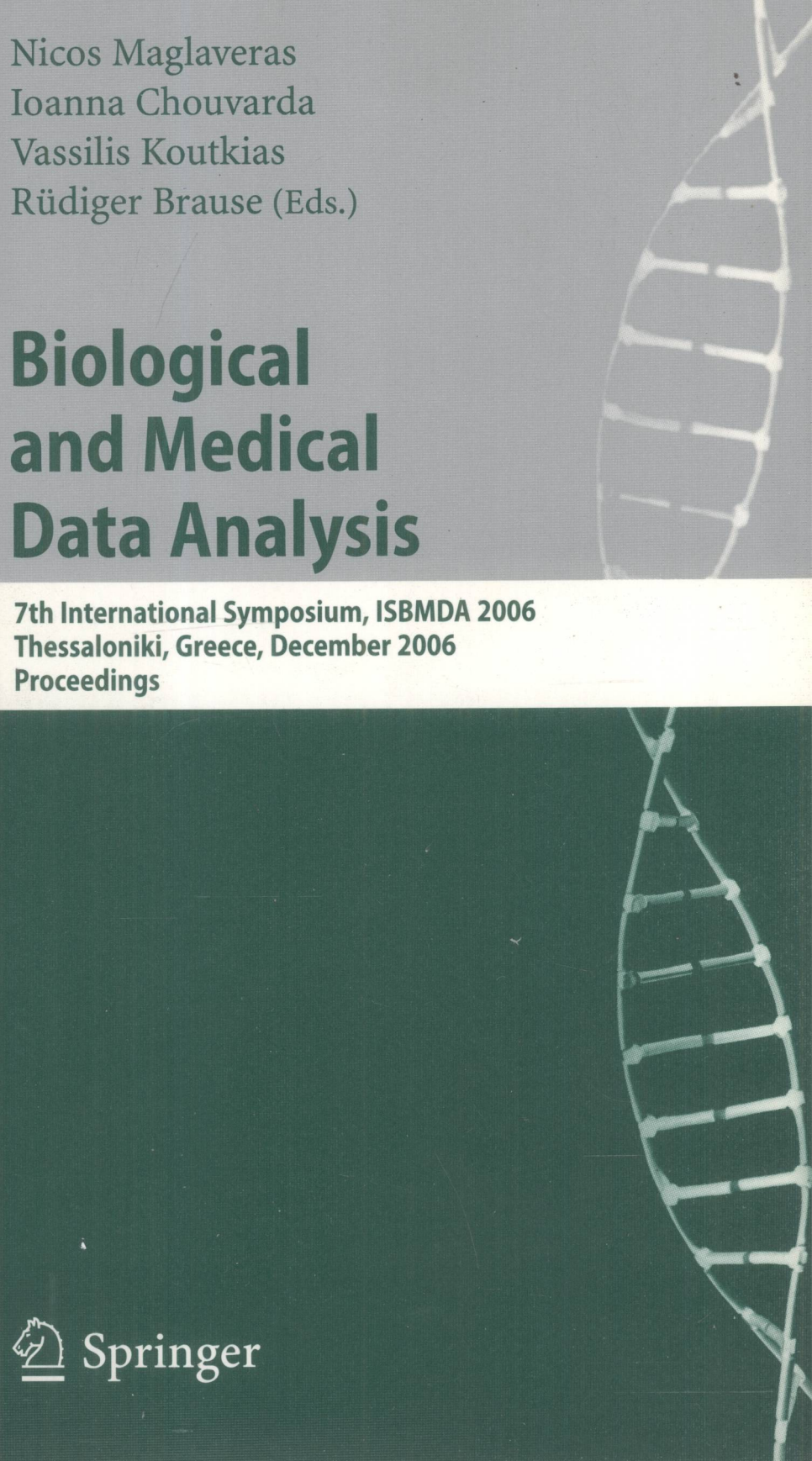
Nicos Maglaveras
Ioanna Chouvarda
Vassilis Koutkias
Rüdiger Brause (Eds.)

LNBI 4345

Biological and Medical Data Analysis

**7th International Symposium, ISBMDA 2006
Thessaloniki, Greece, December 2006
Proceedings**



Springer

Nicos Maglaveras Ioanna Chouvarda
Vassilis Koutkias Rüdiger Brause (Eds.)

Biological and Medical Data Analysis

7th International Symposium, ISBMDA 2006
Thessaloniki, Greece, December 7-8, 2006
Proceedings

Springer

Series Editors

Sorin Istrail, Brown University, Providence, RI, USA
Pavel Pevzner, University of California, San Diego, CA, USA
Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

Nicos Maglaveras
Ioanna Chouvarda
Vassilis Koutkias
Aristotle University
The Medical School
Lab. of Medical Informatics – Box 323
54124 Thessaloniki, Greece
E-mail: {nicmag,ioanna,bikout}@med.auth.gr

Rüdiger Brause
J.W. Goethe-University
Department of Computer Science and Mathematics
Institute for Informatics
Robert-Mayer Str. 11-15, 60054 Frankfurt, Germany
E-mail: rbrause@informatik.uni-frankfurt.de

Library of Congress Control Number: 2006937536

CR Subject Classification (1998): H.2.8, I.2, H.3, G.3, I.5.1, I.4, J.3, F.1

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-68063-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-68063-5 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11946465 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 4345

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Preface

The area of medical and biological data analysis has undergone numerous transformations in recent years. On the one hand, the Human Genome Project has triggered an unprecedented growth in gathering new types of data from the molecular level, which has in turn raised the need for data management and processing and led to the exponential growth of the bioinformatics area. On the other hand, new bits of information coming from molecular data have started filling some long-standing gaps of knowledge, complementing the huge amount of phenotypic data and relevant medical analysis. Thus, bioinformatics, medical informatics and biomedical engineering disciplines, contributing to the vertical integration of knowledge and data, using information technology platforms and enabling technologies such as micro and nano sensors, seem to converge and explore ways to integrate the competencies residing in each field.

ISBMDA has formed a platform, enabling the presentation and integration of new research results adding huge value to this scientific endeavour. This new research information stems from molecular data and phenotypically-oriented medical data analysis such as biosignal or bioimage analysis, novel decision support systems based on new combinations of data, information and extracted knowledge and, innovative information technology solutions enabling the integration of knowledge from the molecules up to the organs and the phenotype. Thus, the 7^{th} ISBMDA was a place for all the above mentioned competencies to come together, and for the discussion and synthesis of new approaches to our understanding of the human organism function on a multiscale level.

We would like to express our gratitude to all the authors who submitted their work to the symposium and gave the Technical Program Committee the opportunity to prepare a symposium of outstanding quality. This year we received 91 contributions and following a rigorous review procedure 44 contributions were selected for presentation at the symposium. The form of the 44 presentations selected was either oral (28 oral presentations) or poster (16 poster presentations). We would finally like to thank the Technical Program Committee and the reviewers who helped, for the preparation of an excellent program for the symposium.

December 2006

Nicos Maglaveras
Ioanna Chouvarda
Vassilis Koutkias
Rüdiger Brause

Organization

Symposium Chair

N. Maglaveras, Aristotle Univ. of Thessaloniki, Greece

Scientific Committee Coordinators

V. Maojo, Univ. Politécnica de Madrid, Spain
F. Martín-Sánchez, Institute of Health Carlos III, Spain
A. Sousa Pereira, Univ. Aveiro, Portugal

Steering Committee

R. Brause, J.W. Goethe Univ., Germany
I. Chouvarda, Aristotle Univ. of Thessaloniki, Greece
V. Koutkias, Aristotle Univ. of Thessaloniki, Greece
A. Malousi, Aristotle Univ. of Thessaloniki, Greece
A. Kokkinaki, Aristotle Univ. of Thessaloniki, Greece

Scientific Committee

A. Babic, Univ. Linkoping, Sweden
R. Baud, Univ. Hospital of Geneva, Switzerland
V. Breton, Univ. Clermont Ferrand, France
J. Carazo, Univ. Autonoma de Madrid, Spain
A. Carvalho, Univ. São Paulo, Brazil
P. Cinquin, Univ. Grenoble, France
W. Dubitzky, Univ. Ulster, UK
M. Dugas, Univ. Munich, Germany
P. Ghazal, Univ. Edinburgh, UK
R. Guthke, Hans-Knoell Institut, Germany
O. Kohlbacher, Univ. Tübingen, Germany
C. Kulikowski, Rutgers Univ., USA
P. Larranaga, Univ. Basque Country, Spain
L. Ohno-Machado, Harvard Univ., USA
J. Luis Oliveira, Univ. Aveiro, Portugal
F. Pinciroli, Politecnico di Milano, Italy
D. Pisanelli, ISTC - CNR, Italy

G. Potamias, ICS - FORTH, Greece
M. Santos, Univ. Aveiro, Portugal
F. Sanz, Univ. Pompeu Fabra, Spain
W. Sauerbrei, Univ. Freiburg, Germany
S. Schulz, Univ. Freiburg, Germany
T. Solomonides, Univ. W. England, UK
C. Zamboulis, Aristotle Univ. of Thessaloniki, Greece
B. Zupan, Univ. Ljubljana, Slovenia
J. Zvárová, Univ. Charles, Czech Republic

Special Reviewers

P.D. Bamidis, Aristotle Univ. of Thessaloniki, Greece
A. Astaras, Aristotle Univ. of Thessaloniki, Greece
S. Kouidou, Aristotle Univ. of Thessaloniki, Greece
A. Malousi, Aristotle Univ. of Thessaloniki, Greece
A. Bezerianos, University of Patras, Greece
D. Perez, Univ. Politécnica de Madrid, Spain
M. Carcia-Remesal, Univ. Politécnica de Madrid, Spain
G. Calle, Univ. Politécnica de Madrid, Spain
J. Crespo, Univ. Politécnica de Madrid, Spain
L. Martin, Univ. Politécnica de Madrid, Spain
D. Manrique, Univ. Politécnica de Madrid, Spain

Lecture Notes in Bioinformatics

Vol. 4345: N. Maglaveras, I. Chouvarda, V. Koutkias, R. Brause (Eds.), Biological and Medical Data Analysis. XIII, 496 pages. 2006.

Vol. 4230: C. Priami, A. Ingólfsdóttir, B. Mishra, H.R. Nielson (Eds.), Transactions on Computational Systems Biology VII. VII, 185 pages. 2006.

Vol. 4220: C. Priami, G. Plotkin (Eds.), Transactions on Computational Systems Biology VI. VII, 247 pages. 2006.

Vol. 4216: M.R. Berthold, R. Glen, I. Fischer (Eds.), Computational Life Sciences II. XIII, 269 pages. 2006.

Vol. 4210: C. Priami (Ed.), Computational Methods in Systems Biology. X, 323 pages. 2006.

Vol. 4205: G. Bourque, N. El-Mabrouk (Eds.), Comparative Genomics. X, 231 pages. 2006.

Vol. 4175: P. Bücher, B.M.E. Moret (Eds.), Algorithms in Bioinformatics. XII, 402 pages. 2006.

Vol. 4146: J.C. Rajapakse, L. Wong, R. Acharya (Eds.), Pattern Recognition in Bioinformatics. XIV, 186 pages. 2006.

Vol. 4115: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence and Bioinformatics, Part III. XXI, 803 pages. 2006.

Vol. 4075: U. Leser, F. Naumann, B. Eckman (Eds.), Data Integration in the Life Sciences. XI, 298 pages. 2006.

Vol. 4070: C. Priami, X. Hu, Y. Pan, T.Y. Lin (Eds.), Transactions on Computational Systems Biology V. IX, 129 pages. 2006.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006.

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006.

Vol. 3886: E.G. Bremer, J. Hakenberg, E.-H.(S.) Han, D. Berrar, W. Dubitzky (Eds.), Knowledge Discovery in Life Science Literature. XIV, 147 pages. 2006.

Vol. 3745: J.L. Oliveira, V. Maojo, F. Martín-Sánchez, A.S. Pereira (Eds.), Biological and Medical Data Analysis. XII, 422 pages. 2005.

Vol. 3737: C. Priami, E. Merelli, P. Gonzalez, A. Omicini (Eds.), Transactions on Computational Systems Biology III. VII, 169 pages. 2005.

Vol. 3695: M.R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005.

Vol. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005.

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005.

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005.

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VII, 133 pages. 2005.

Vol. 3380: C. Priami (Ed.), Transactions on Computational Systems Biology I. IX, 111 pages. 2005.

Vol. 3370: A. Konagaya, K. Satou (Eds.), Grid Computing in Life Science. X, 188 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VII, 115 pages. 2005.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004.

Vol. 3082: V. Danos, V. Schachter (Eds.), Computational Methods in Systems Biology. IX, 280 pages. 2005.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004.

Vol. 2812: G. Benson, R.D.M. Page (Eds.), Algorithms in Bioinformatics. X, 528 pages. 2003.

Vol. 2666: C. Guerra, S. Istrail (Eds.), Mathematical Methods for Protein Structure Analysis and Design. XI, 157 pages. 2003.

Printing: Mercedes-Druck, Berlin
Binding: Stein + Lehmann, Berlin

¥547.00元

Table of Contents

Bioinformatics: Functional Genomics

HLA and HIV Infection Progression: Application of the Minimum Description Length Principle to Statistical Genetics 1
Peter T. Hraber, Bette T. Korber, Steven Wolinsky, Henry A. Erlich, Elizabeth A. Trachtenberg, and Thomas B. Kepler

Visualization of Functional Aspects of microRNA Regulatory Networks Using the Gene Ontology 13
Alkiviadis Symeonidis, Ioannis G. Tollis, and Martin Reczko

A Novel Method for Classifying Subfamilies and Sub-subfamilies of G-Protein Coupled Receptors 25
Majid Beigi and Andreas Zell

Integration Analysis of Diverse Genomic Data Using Multi-clustering Results 37
Hye-Sung Yoon, Sang-Ho Lee, Sung-Bum Cho, and Ju Han Kim

Bioinformatics: Sequence and Structure Analysis

Effectivity of Internal Validation Techniques for Gene Clustering 49
Chunmei Yang, Baikun Wan, and Xiaofeng Gao

Intrinsic Splicing Profile of Human Genes Undergoing Simple Cassette Exon Events 60
Andigoni Malousi, Vassilis Koutkias, Sofia Kouidou, and Nicos Maglaveras

Generalization Rules for Binarized Descriptors 72
Jürgen Paetz

Application of Combining Classifiers Using Dynamic Weights to the Protein Secondary Structure Prediction – Comparative Analysis of Fusion Methods 83
Tomasz Woloszynski and Marek Kurzynski

A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites 92
George Tzanis, Christos Berberidis, and Ioannis Vlahavas

SPSO: Synthetic Protein Sequence Oversampling for Imbalanced Protein Data and Remote Homology Detection 104
Majid Beigi and Andreas Zell

Biomedical Models

Markov Modeling of Conformational Kinetics of Cardiac Ion Channel Proteins 116
Chong Wang, Antje Krause, Chris Nugent, and Werner Dubitzky

Insulin Sensitivity and Plasma Glucose Appearance Profile by Oral Minimal Model in Normotensive and Normoglycemic Humans 128
Roberto Burattini, Fabrizio Casagrande, Francesco Di Nardo, Massimo Boemi, and Pierpaolo Morosini

Dynamic Model of Amino Acid and Carbohydrate Metabolism in Primary Human Liver Cells 137
Reinhard Guthke, Wolfgang Schmidt-Heck, Gesine Pless, Rolf Gebhardt, Michael Pfaff, Joerg C. Gerlach, and Katrin Zeilinger

The Probabilities Mixture Model for Clustering Flow-Cytometric Data: An Application to Gating Lymphocytes in Peripheral Blood 150
John Lakoumentas, John Drakos, Marina Karakantza, Nicolaos Zoumbos, George Nikiforidis, and George Sakellaropoulos

Integrative Mathematical Modeling for Analysis of Microcirculatory Function 161
Adam Kapela, Anastasios Bezerianos, and Nikolaos Tsoukias

Searching and Visualizing Brain Networks in Schizophrenia 172
Theofanis Oikonomou, Vangelis Sakkalis, Ioannis G. Tollis, and Sifis Micheloyannis

Databases and Grids

TRENCADIS – A Grid Architecture for Creating Virtual Repositories of DICOM Objects in an OGSA-Based Ontological Framework 183
Ignacio Blanquer, Vicente Hernandez, and Damià Segrelles

Minimizing Data Size for Efficient Data Reuse in Grid-Enabled Medical Applications 195
Fumihiko Ino, Katsunori Matsuo, Yasuharu Mizutani, and Kenichi Hagihara

Thinking Precedes Action: Using Software Engineering for the Development of a Terminology Database to Improve Access to Biomedical Documentation 207
Antonio Vaquero, Fernando Sáenz, Francisco Álvarez, and Manuel de Buenaga

Grid-Based Knowledge Discovery in Clinico-Genomic Data 219
Michael May, George Potamias, and Stefan Rüping

A Prospective Study on the Integration of Microarray Data in HIS/EPR 231
Daniel F. Polónia, Joel Arrais, and José Luis Oliveira

Web Services Interface to Run Protein Sequence Tools on Grid, Testcase of Protein Sequence Alignment 240
Christophe Blanchet, Christophe Combet, Vladimir Daric, and Gilbert Deléage

Semantics and Information Modelling

Integrating Clinical and Genomic Information Through the PrognoChip Mediator 250
Anastasia Analyti, Haridimos Kondylakis, Dimitris Manakanatas, Manos Kalaitzakis, Dimitris Plexousakis, and George Potamias

OntoDataClean: Ontology-Based Integration and Preprocessing of Distributed Data 262
David Perez-Rey, Alberto Anguita, and Jose Crespo

Language Modelling for the Needs of OCR of Medical Texts 273
Maciej Piasecki and Grzegorz Godlewski

Biomedical Signal Processing – Time Series Analysis

The Use of Multivariate Autoregressive Modelling for Analyzing Dynamical Physiological Responses of Individual Critically Ill Patients 285
Kristien Van Loon, Jean-Marie Aerts, Geert Meyfroidt, Greta Van den Berghe, and Daniel Berckmans

Time Series Feature Evaluation in Discriminating Preictal EEG States 298
Dimitris Kugiumtzis, Angeliki Papana, Alkiviadis Tsimpiris, Ioannis Vlachos, and Pål G. Larsson

Symbol Extraction Method and Symbolic Distance for Analysing Medical Time Series 311
Fernando Alonso, Loïc Martínez, Aurora Pérez, Agustín Santamaría, and Juan Pedro Valente

A Wavelet Tool to Discriminate Imagery Versus Actual Finger Movements Towards a Brain–Computer Interface 323
Maria L. Stavrinou, Liviu Moraru, Polyxeni Pelekouda, Vasileios Kokkinos, and Anastasios Bezerianos

Biomedical Image Analysis and Visualisation Techniques

A Fully Bayesian Two-Stage Model for Detecting Brain Activity in fMRI 334
Alicia Quirós, Raquel Montes Diez, and Juan A. Hernández

A Novel Algorithm for Segmentation of Lung Images 346
Aamir Saeed Malik and Tae-Sun Choi

An Evaluation of Image Compression Algorithms for Colour Retinal Images 358
Gerald Schaefer and Roman Starosolski

An Automated Model for Rapid and Reliable Segmentation of Intravascular Ultrasound Images 368
Eirini Parissi, Yiannis Kompatsiaris, Yiannis S. Chatzizisis, Vassilis Koutkias, Nicos Maglaveras, M.G. Strintzis, and George D. Giannoglou

Biomedical Data Analysis and Interpretation

Supervised Neuro-fuzzy Clustering for Life Science Applications 378
Jürgen Paetz

Study on Preprocessing and Classifying Mass Spectral Raw Data Concerning Human Normal and Disease Cases 390
Xenofon E. Floros, George M. Spyrou, Konstantinos N. Vougas, George T. Tsangaris, and Konstantina S. Nikita

Non-repetitive DNA Sequence Compression Using Memoization 402
K.G. Srinivasa, M. Jagadish, K.R. Venugopal, and L.M. Patnaik

Application of Rough Sets Theory to the Sequential Diagnosis 413
Andrzej Zolnierek

Data Integration in Multi-dimensional Data Sets: Informational Asymmetry in the Valid Correlation of Subdivided Samples 423
Qing T. Zeng, Juan Pablo Pratt, Jane Pak, Eun-Young Kim, Dino Ravnic, Harold Huss, and Steven J. Mentzer

Decision Support Systems and Diagnostic Tools

Two-Stage Classifier for Diagnosis of Hypertension Type 433
Michal Wozniak

Handwriting Analysis for Diagnosis and Prognosis of Parkinson's Disease 441
Atilla Ünlü, Rüdiger Brause, and Karsten Krakow

A Decision Support System for the Automatic Assessment of Hip Osteoarthritis Severity by Hip Joint Space Contour Spectral Analysis ... 451
Ioannis Boniatis, Dionisis Cavouras, Lena Costaridou, Ioannis Kalatzis, Elias Panagiotopoulos, and George Panayiotakis

Modeling for Missing Tissue Compensator Fabrication Using RFID Tag in U-Health ... 463
O-Hoon Choi, Jung-Eun Lim, Hong-Seok Na, and Doo-Kwon Baik

The Effect of User Factors on Consumer Familiarity with Health Terms: Using Gender as a Proxy for Background Knowledge About Gender-Specific Illnesses ... 472
Alla Keselman, Lisa Massengale, Long Ngo, Allen Browne, and Qing Zeng

ICT for Patient Safety: Towards a European Research Roadmap ... 482
Veli N. Stroetmann, Daniel Spichtinger, Karl A. Stroetmann, and Jean Pierre Thierry

Author Index ... 495

HLA and HIV Infection Progression: Application of the Minimum Description Length Principle to Statistical Genetics

Peter T. Hraber[1,2], Bette T. Korber[1,2], Steven Wolinsky[3], Henry A. Erlich[4], Elizabeth A. Trachtenberg[5], and Thomas B. Kepler[6]

[1] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501 USA
[2] Los Alamos National Laboratory, Los Alamos NM 87545 USA
[3] Feinberg School of Medicine, Northwestern University, Chicago IL 60611 USA
[4] Roche Molecular Systems, 1145 Atlantic Ave, Alameda CA 94501 USA
[5] Children's Hospital Oakland Research Institute, Oakland CA 94609 USA
[6] Department of Biostatistics and Bioinformatics, Duke University Medical Center, Duke University, Durham NC 27708 USA

Abstract. The minimum description length (MDL) principle was developed in the context of computational complexity and coding theory. It states that the best model to account for some data minimizes the sum of the lengths, in bits, of the descriptions of the model and the data as encoded via the model. The MDL principle gives a criterion for parameter selection, by using the description length as a test statistic. Class I HLA genes play a major role in the immune response to HIV, and are known to be associated with rates of progression to AIDS. However, these genes are extremely polymorphic, making it difficult to associate alleles with disease outcome, given statistical issues of multiple testing. Application of the MDL principle to immunogenetic data from a longitudinal cohort study (Chicago MACS) enables classification of alleles associated with plasma HIV RNA abundance, an indicator of infection progression. Variation in progression is strongly associated with HLA-B. Allele associations with viral levels support and extend previous studies. In particular, individuals without *B58s* supertype alleles average viral RNA levels 3.6 times greater than individuals with them. Mechanisms for these associations include variation in epitope specificity and selection that favors rare alleles.

1 Introduction

Progression of HIV infection is characterized by three phases: acute, or early, chronic, and AIDS, the final phase of infection preceeding death [1]. The chronic phase is variable in duration, lasting ten years on average, but varying from two to twenty years. A good predictor of the duration of the chronic phase is the viral RNA level during chronic infection, with higher levels consistently associated with more rapid progression than lower levels [2]. A major challenge for treating HIV and developing effective vaccination strategies is to understand what causes variation in plasma viral RNA levels, and hence to infection progression.

N. Maglaveras et al. (Eds.): ISBMDA 2006, LNBI 4345, pp. 1–12, 2006.

The cell-mediated immune response identifies and eliminates infected cells from an individual. A central role in this response is played by the major histocompatibility complex (MHC), in humans, also known as human leukocyte antigens (HLA). Two classes of HLA genes code for codominately expressed cell-surface glycoproteins, and present processed peptide to circulating T cells, which discriminate between self (uninfected) and non-self (infected) cells [3,4].

Class I HLA molecules are expressed on all nucleated cells except germ cells. In infected cells, they bind and present antigenic peptide fragments to T-cell receptors on CD8+ T lymphocytes, which are usually cytotoxic and cause lysis of the infected cell. Class II HLA molecules are expressed on immunogenetically reactive cells, such as dendritic cells, B cells, macrophages, and activated T cells. They present antigen peptide fragments to T-cell receptors on CD4+ T-lymphocytes and the interaction results in release of cytokines that stimulate the immune response.

Human HLA loci are among the most diverse known [5,6]. This diversity provides a repertoire to recognize evolving antigens [6,7]. Previous studies of associations between HLA alleles and variation in progression of HIV-1 infection have established that within-host HLA diversity helps to inhibit viral infection, by associating degrees of heterozygosity with rates of HIV disease progression [8]. Thus, homozygous individuals, particularly at the HLA-B locus, suffer a greater rate of progression than do heterozygotes [8,9]. Identifying which alleles are associated with variation in rates of infection progression has been difficult, due in part to the compounding of error rates incurred when testing many alternative hypotheses, and published results do not always agree [10,11].

This study demonstrates the use of an information-based criterion for statistical inference. Its approach to multiple testing differs from that of standard analytic techniques, and provides the ability to resolve associations between variation in HIV RNA abundance and variation in HLA alleles.

As an application of computational complexity and optimal coding theory to statistical inference, the minimum description length (MDL) principle states that the best statistical model, or hypothesis, to account for some observed data is the model that minimizes the sum of the number of bits required to describe both the model and the data encoded via the model [12,13,14]. It is a model-selection criterion that balances the need for parsimony and fidelity, by penalizing equally for the information required to specify the model and the information required to encode the residual error.

2 Methods

The analyses detailed below apply the MDL principle to the problem of partitioning individuals into groups having similar HIV RNA levels, based on HLA alleles present in each case.

Chicago MACS HLA & HIV Data. The Chicago Multicenter AIDS Cohort Study (MACS) provided an opportunity to analyze a detailed, long-term, longitudinal set of clinical HIV/HLA data [10]. Each participant provided informed

consent in writing. Of 564 HIV-positive cases sampled in the Chicago MACS, 479 provided information about both the rate of disease progression and HLA genetic background. Progression was indicated by the quasi-stationary "set-point" viral RNA level during chronic infection. Immunogenetic background was obtained by genotyping HLA alleles from class I (HLA-A, -B, and -C) and class II (HLA-DRB1, -DQB1, and -DPB1) loci.

Viral RNA set-point levels were determined after acute infection and prior to any therapeutic intervention or the onset of AIDS, as defined by the presence of an opportunistic infection or CD4+ T-cell count below 200 per ml of plasma. Because the assay has a detection threshold of 300 copies of virus per ml [10], maximum-likelihood estimators were adjusted to avoid biased estimates of population parameters from a truncated, or censored, sample distribution [15]. Viral RNA levels were log-transformed for better approximations to normality.

High-resolution class I and II HLA genotyping [10] provided four-digit allele designations, though analyses were generally performed using two-digit allele designations because of the resulting reduction of allelic diversity and increased number of samples per allele. Because of the potential for results to be confounded by an effect associated with an individual's ethnicity or revised sampling protocol, two separate analyses were performed, one using data from the entire cohort, and another using only data from Caucasian individuals. Sample numbers were too small to study other subgroups independently.

HLA supertypes group class I alleles by their peptide-binding anchor motifs [16]. Assignment of four-digit allele designations to functionally related groups of supertypes at HLA-A and -B loci facilitated further analysis. Where they could be determined, HLA-A and HLA-B supertypes were assigned from four-digit allele designations [10]. As with two-digit allele designations for each locus, HLA-A and -B supertypes were assessed for association with viral RNA levels. Cases having other alleles were withheld from classification and subsequent analysis.

A description length analysis determined whether HIV RNA levels were non-trivially associated with alleles at any HLA locus.

Description Lengths. The challenge of data classification is to find the best partition, such that observations within a group are well-described as independent draws from a single population, but differences in population distributions exist between groups. Whether the data are better represented as two groups, or more, than as one depends on the description lengths that result.

We use the family of Gaussian distributions to model viral RNA levels. While the MDL strategy can be applied using any probabilistic model, a log-normal distribution is a good choice for the observed plasma viral RNA values. First, the description length of the model and of the data given the model is calculated as described below, grouping all of the observations into one normal distribution, L_1. Next, the data are broken into two partitions, L_2, and the log-RNA values associated with HLA alleles are partitioned to minimize the description length given the constraint that two Gaussian distributions, each having their own mean and variance, are used to model the data.