Elena Marchiori
Jason H. Moore
Jagath C. Rajapakse (Eds.)

LNCS 4447

Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics

5th European Conference, EvoBIO 2007
Valencia, Spain, April 2007
Proceedings

Springer

Elena Marchiori Jason H. Moore
Jagath C. Rajapakse (Eds.)

Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics

5th European Conference, EvoBIO 2007
Valencia, Spain, April 11-13, 2007
Proceedings

Springer

E2007003202

Volume Editors

Elena Marchiori
VU University of Amsterdam, IBIVU
Department of Computer Science
de Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
E-mail: elena@cs.vu.nl

Jason H. Moore
Dartmouth-Hitchcock Medical Center
Computational Genetics Laboratory
706 Rubin Building, HB 7937, One Medical Center Dr., Lebanon, NH 03756, USA
E-mail: jason.h.moore@dartmouth.edu

Jagath C. Rajapakse
Nanyang Technological University
School of Computer Engineering
Blk N4-2a05, 50 Nanyang Avenue, Singapore 639798
E-mail: asjagath@ntu.edu.sg

Cover illustration: Morphogenesis series #12 by Jon McCormack, 2006

Library of Congress Control Number: 2007923724

CR Subject Classification (1998): D.1, F.1-2, J.3, I.5, I.2

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-71782-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-71782-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12044597 06/3180 5 4 3 2 1 0

Lecture Notes in Computer Science 4447

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Lecture Notes in Computer Science

For information about Vols. 1–4336

please contact your bookseller or Springer

Vol. 4447: E. Marchiori, J.H. Moore, J.C. Rajapakse (Eds.), Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. XI, 302 pages. 2007.

Vol. 4446: C. Cotta, J. van Hemert (Eds.), Evolutionary Computation in Combinatorial Optimization. XII, 241 pages. 2007.

Vol. 4445: M. Ebner, M. O'Neill, A. Ekárt, L. Vanneschi, A.I. Esparcia-Alcázar (Eds.), Genetic Programming. XI, 382 pages. 2007.

Vol. 4444: T. Reps, M. Sagiv, J. Bauer (Eds.), Program Analysis and Compilation, Theory and Practice. X, 361 pages. 2007.

Vol. 4443: R. Kotagiri, P.R. Krishna, M.K. Mohania, E. Nantajeewarawat (Eds.), Advances in Databases: Concepts, Systems and Applications. XXI, 1126 pages. 2007.

Vol. 4430: C.C. Yang, D. Zeng, M. Chau, K. Chang, Q. Yang, X. Cheng, J. Wang, F.-Y. Wang, H. Chen (Eds.), Intelligence and Security Informatics. XII, 330 pages. 2007.

Vol. 4429: R. Lu, J.H. Siekmann, C. Ullrich (Eds.), Cognitive Systems. X, 161 pages. 2007. (Sublibrary LNAI).

Vol. 4427: S. Uhlig, K. Papagiannaki, O. Bonaventure (Eds.), Passive and Active Network Measurement. XI, 274 pages. 2007.

Vol. 4425: G. Amati, C. Carpineto, G. Romano (Eds.), Advances in Information Retrieval. XIX, 759 pages. 2007.

Vol. 4424: O. Grumberg, M. Huth (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XX, 738 pages. 2007.

Vol. 4423: H. Seidl (Ed.), Foundations of Software Science and Computational Structures. XVI, 379 pages. 2007.

Vol. 4422: M.B. Dwyer, A. Lopes (Eds.), Fundamental Approaches to Software Engineering. XV, 440 pages. 2007.

Vol. 4421: R. De Nicola (Ed.), Programming Languages and Systems. XVII, 538 pages. 2007.

Vol. 4420: S. Krishnamurthi, M. Odersky (Eds.), Compiler Construction. XIV, 233 pages. 2007.

Vol. 4419: P.C. Diniz, E. Marques, K. Bertels, M.M. Fernandes, J.M.P. Cardoso (Eds.), Reconfigurable Computing: Architectures, Tools and Applications. XIV, 391 pages. 2007.

Vol. 4418: A. Gagalowicz, W. Philips (Eds.), Computer Vision/Computer Graphics Collaboration Techniques. XV, 620 pages. 2007.

Vol. 4416: A. Bemporad, A. Bicchi, G. Buttazzo (Eds.), Hybrid Systems: Computation and Control. XVII, 797 pages. 2007.

Vol. 4415: P. Lukowicz, L. Thiele, G. Tröster (Eds.), Architecture of Computing Systems - ARCS 2007. X, 297 pages. 2007.

Vol. 4414: S. Hochreiter, R. Wagner (Eds.), Bioinformatics Research and Development. XVI, 482 pages. 2007. (Sublibrary LNBI).

Vol. 4410: A. Branco (Ed.), Anaphora: Analysis, Algorithms and Applications. X, 191 pages. 2007. (Sublibrary LNAI).

Vol. 4407: G. Puebla (Ed.), Logic-Based Program Synthesis and Transformation. VIII, 237 pages. 2007.

Vol. 4405: L. Padgham, F. Zambonelli (Eds.), Agent-Oriented Software Engineering VII. XII, 225 pages. 2007.

Vol. 4403: S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, T. Murata (Eds.), Evolutionary Multi-Criterion Optimization. XIX, 954 pages. 2007.

Vol. 4400: J.F. Peters, A. Skowron, V.W. Marek, E. Orłowska, R. Slowinski, W. Ziarko (Eds.), Transactions on Rough Sets VII, Part II. X, 381 pages. 2007.

Vol. 4399: X. Llorà, T. Kovacs, K. Takadama, P.L. Lanzi, S.W. Wilson, W. Stolzmann (Eds.), Learning Classifier Systems. XII, 345 pages. 2007. (Sublibrary LNAI).

Vol. 4398: S. Marchand-Maillet, E. Bruno, A. Nürnberger, M. Detyniecki (Eds.), Adaptive Multimedia Retrieval: User, Context, and Feedback. XI, 269 pages. 2007.

Vol. 4397: C. Stephanidis, M. Pieper (Eds.), Universal Access in Ambient Intelligence Environments. XV, 467 pages. 2007.

Vol. 4396: J. García-Vidal, L. Cerdà-Alabern (Eds.), Wireless Systems and Mobility in Next Generation Internet. IX, 271 pages. 2007.

Vol. 4395: M. Daydé, J.M.L.M. Palma, Á.L.G.A. Coutinho, E. Pacitti, J.C. Lopes (Eds.), High Performance Computing for Computational Science - VECPAR 2006. XXIV, 721 pages. 2007.

Vol. 4394: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. XVI, 648 pages. 2007.

Vol. 4393: W. Thomas, P. Weil (Eds.), STACS 2007. XVIII, 708 pages. 2007.

Vol. 4392: S.P. Vadhan (Ed.), Theory of Cryptography. XI, 595 pages. 2007.

Vol. 4391: Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.), Progress in Nonlinear Speech Processing. XII, 269 pages. 2007.

Vol. 4390: S.O. Kuznetsov, S. Schmidt (Eds.), Formal Concept Analysis. X, 329 pages. 2007. (Sublibrary LNAI).

Vol. 4389: D. Weyns, H.V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems III. X, 273 pages. 2007. (Sublibrary LNAI).

Vol. 4385: K. Coninx, K. Luyten, K.A. Schneider (Eds.), Task Models and Diagrams for Users Interface Design. XI, 355 pages. 2007.

Vol. 4384: T. Washio, K. Satoh, H. Takeda, A. Inokuchi (Eds.), New Frontiers in Artificial Intelligence. IX, 401 pages. 2007. (Sublibrary LNAI).

Vol. 4383: E. Bin, A. Ziv, S. Ur (Eds.), Hardware and Software, Verification and Testing. XII, 235 pages. 2007.

Vol. 4381: J. Akiyama, W.Y.C. Chen, M. Kano, X. Li, Q. Yu (Eds.), Discrete Geometry, Combinatorics and Graph Theory. XI, 289 pages. 2007.

Vol. 4380: S. Spaccapietra, P. Atzeni, F. Fages, M.-S. Hacid, M. Kifer, J. Mylopoulos, B. Pernici, P. Shvaiko, J. Trujillo, I. Zaihrayeu (Eds.), Journal on Data Semantics VIII. XV, 219 pages. 2007.

Vol. 4378: I. Virbitskaite, A. Voronkov (Eds.), Perspectives of Systems Informatics. XIV, 496 pages. 2007.

Vol. 4377: M. Abe (Ed.), Topics in Cryptology – CT-RSA 2007. XI, 403 pages. 2006.

Vol. 4376: E. Frachtenberg, U. Schwiegelshohn (Eds.), Job Scheduling Strategies for Parallel Processing. VII, 257 pages. 2007.

Vol. 4374: J.F. Peters, A. Skowron, I. Düntsch, J. Grzymała-Busse, E. Orłowska, L. Polkowski (Eds.), Transactions on Rough Sets VI, Part I. XII, 499 pages. 2007.

Vol. 4373: K. Langendoen, T. Voigt (Eds.), Wireless Sensor Networks. XIII, 358 pages. 2007.

Vol. 4372: M. Kaufmann, D. Wagner (Eds.), Graph Drawing. XIV, 454 pages. 2007.

Vol. 4371: K. Inoue, K. Satoh, F. Toni (Eds.), Computational Logic in Multi-Agent Systems. X, 315 pages. 2007. (Sublibrary LNAI).

Vol. 4370: P.P Lévy, B. Le Grand, F. Poulet, M. Soto, L. Darago, L. Toubiana, J.-F. Vibert (Eds.), Pixelization Paradigm. XV, 279 pages. 2007.

Vol. 4369: M. Umeda, A. Wolf, O. Bartenstein, U. Geske, D. Seipel, O. Takata (Eds.), Declarative Programming for Knowledge Management. X, 229 pages. 2006. (Sublibrary LNAI).

Vol. 4368: T. Erlebach, C. Kaklamanis (Eds.), Approximation and Online Algorithms. X, 345 pages. 2007.

Vol. 4367: K. De Bosschere, D. Kaeli, P. Stenström, D. Whalley, T. Ungerer (Eds.), High Performance Embedded Architectures and Compilers. XI, 307 pages. 2007.

Vol. 4366: K. Tuyls, R. Westra, Y. Saeys, A. Nowé (Eds.), Knowledge Discovery and Emergent Complexity in Bioinformatics. IX, 183 pages. 2007. (Sublibrary LNBI).

Vol. 4364: T. Kühne (Ed.), Models in Software Engineering. XI, 332 pages. 2007.

Vol. 4362: J. van Leeuwen, G.F. Italiano, W. van der Hoek, C. Meinel, H. Sack, F. Plášil (Eds.), SOFSEM 2007: Theory and Practice of Computer Science. XXI, 937 pages. 2007.

Vol. 4361: H.J. Hoogeboom, G. Păun, G. Rozenberg, A. Salomaa (Eds.), Membrane Computing. IX, 555 pages. 2006.

Vol. 4360: W. Dubitzky, A. Schuster, P.M.A. Sloot, M. Schroeder, M. Romberg (Eds.), Distributed, High-Performance and Grid Computing in Computational Biology. X, 192 pages. 2007. (Sublibrary LNBI).

Vol. 4358: R. Vidal, A. Heyden, Y. Ma (Eds.), Dynamical Vision. IX, 329 pages. 2007.

Vol. 4357: L. Buttyán, V. Gligor, D. Westhoff (Eds.), Security and Privacy in Ad-Hoc and Sensor Networks. X, 193 pages. 2006.

Vol. 4355: J. Julliand, O. Kouchnarenko (Eds.), B 2007: Formal Specification and Development in B. XIII, 293 pages. 2006.

Vol. 4354: M. Hanus (Ed.), Practical Aspects of Declarative Languages. X, 335 pages. 2006.

Vol. 4353: T. Schwentick, D. Suciu (Eds.), Database Theory – ICDT 2007. XI, 419 pages. 2006.

Vol. 4352: T.-J. Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua, L.-T. Chia (Eds.), Advances in Multimedia Modeling, Part II. XVIII, 743 pages. 2006.

Vol. 4351: T.-J. Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua, L.-T. Chia (Eds.), Advances in Multimedia Modeling, Part I. XIX, 797 pages. 2006.

Vol. 4349: B. Cook, A. Podelski (Eds.), Verification, Model Checking, and Abstract Interpretation. XI, 395 pages. 2007.

Vol. 4348: S.T. Taft, R.A. Duff, R.L. Brukardt, E. Ploedereder, P. Leroy (Eds.), Ada 2005 Reference Manual. XXII, 765 pages. 2006.

Vol. 4347: J. Lopez (Ed.), Critical Information Infrastructures Security. X, 286 pages. 2006.

Vol. 4346: L. Brim, B. Haverkort, M. Leucker, J. van de Pol (Eds.), Formal Methods: Applications and Technology. X, 363 pages. 2007.

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

Vol. 4344: V. Gruhn, F. Oquendo (Eds.), Software Architecture. X, 245 pages. 2006.

Vol. 4342: H. de Swart, E. Orłowska, G. Schmidt, M. Roubens (Eds.), Theory and Applications of Relational Structures as Knowledge Instruments II. X, 373 pages. 2006. (Sublibrary LNAI).

Vol. 4341: P.Q. Nguyen (Ed.), Progress in Cryptology - VIETCRYPT 2006. XI, 385 pages. 2006.

Vol. 4340: R. Prodan, T. Fahringer, Grid Computing. XXIII, 317 pages. 2007.

Vol. 4339: E. Ayguadé, G. Baumgartner, J. Ramanujam, P. Sadayappan (Eds.), Languages and Compilers for Parallel Computing. XI, 476 pages. 2006.

Vol. 4338: P. Kalra, S. Peleg (Eds.), Computer Vision, Graphics and Image Processing. XV, 965 pages. 2006.

Vol. 4337: S. Arun-Kumar, N. Garg (Eds.), FSTTCS 2006: Foundations of Software Technology and Theoretical Computer Science. XIII, 430 pages. 2006.

¥516.00元

Preface

The field of bioinformatics has two main objectives: the creation and maintenance of biological databases, and the discovery of knowledge from life sciences data in order to unravel the mysteries of biological function, leading to new drugs and therapies for human disease. Life sciences data come in the form of biological sequences, structures, pathways, or literature. One major aspect of discovering biological knowledge is to search, predict, or model specific patterns present in a given dataset and then to interpret those patterns. Computer science methods such as evolutionary computation, machine learning, and data mining all have a great deal to offer the field of bioinformatics. The goal of the Fifth European Conference on Evolutionary Computation, Machine Learning, and Data Mining in Bioinformatics (EvoBIO 2007) was to bring experts in computer science together with experts in bioinformatics and the biological sciences to explore new and novel methods for solving complex biological problems.

The fifth EvoBIO conference was held in Valencia, Spain during April 11-13, 2007 at the Universidad Politecnica de Valencia. EvioBIO 2007 was held jointly with the Tenth European Conference on Genetic Programming (EuroGP 2007), the Seventh European Conference on Evolutionary Computation in Combinatorial Optimisation (EvoCOP 2007), and the Evo Workshops. Collectively, the conferences and workshops are organized under the name Evo* (`www.evostar.org`).

EvoBIO, held annually as a workshop since 2003, became a conference in 2007 and it is now the premiere European event for those interested in the interface between evolutionary computation, machine learning, data mining, bioinformatics, and computational biology. All papers in this book were presented at EvoBIO 2007 and responded to a call for papers that included topics of interest such as biomarker discovery, cell simulation and modeling, ecological modeling, fluxomics, gene networks, biotechnology, metabolomics, microarray analysis, phylogenetics, protein interactions, proteomics, sequence analysis and alignment, and systems biology. A total of 60 papers were submitted to the conference for double-blind peer-review. Of those, 28 (46.7%) were accepted.

We would first and foremost like to thank all authors who spent time and effort to make important contributions to this book. We would like to thank the members of the Program Committee for their expert evaluation of the submitted papers. Moreover, we would like to thank Jennifer Willies for her tremendous administrative help and coordination, Anna Isabel Esparcia-Alcázar for serving as the Local Chair, Leonardo Vanneschi for serving as Evo* Publicity Chair, Marc Schoenauer and the MyReview team (`http://myreview.lri.fr/`) for the conference management system.

We would also like to acknowledge the following organizations. The Universidad Politécnica de Valencia, Spain for their institutional and financial support, and for providing premises and administrative assistance; the Instituto

Tecnológico de Informática in Valencia, for cooperation and help with local arrangements; the Spanish Ministerio de Educación y Ciencia, for their financial support; and the Centre for Emergent Computing at Napier University in Edinburgh, Scotland for administrative support and event coordination.

Finally, we hope that you will consider contributing to EvoBIO 2008.

February 2007

Elena Marchiori
Jason H. Moore
Jagath C. Rajapakse

Organization

EvoBIO 2007 was organized by Evo* (`www.evostar.org`).

Program Chairs

Elena Marchiori (IBIVU, VU University Amsterdam, The Netherlands)
Jason H. Moore (Dartmouth Medical School in Lebanon, NH,USA)
Jagath C. Rajapakse (Nanyang Technological University, Singapore)

General Chairs

David W. Corne (Heriot-Watt University, Edinburgh, UK)
Elena Marchiori (IBIVU, VU University Amsterdam, The Netherlands)

Steering Committee

David W. Corne (Heriot-Watt University, Edinburgh, UK)
Elena Marchiori (IBIVU, VU University Amsterdam, The Netherlands)
Carlos Cotta (University of Malaga, Spain)
Jason H. Moore (Dartmouth Medical School in Lebanon, NH,USA)
Jagath C. Rajapakse (Nanyang Technological University, Singapore)

Program Committee

Jesus S. Aguilar-Ruiz (Spain)
Francisco J. Azuaje (UK)
Wolfgang Banzhaf (Canada)
Jacek Blazewicz (Poland)
Marius Codrea (The Netherlands)
Dave Corne (UK)
Carlos Cotta (Spain)
Alex Freitas (UK)
Gary Fogel (USA)
James Foster (USA)
Rosalba Giugno (Italy)
Raul Giraldez (Spain)
Jin-Kao Hao (France)
Antoine van Kampen (The Netherlands)
Natalio Krasnogor (UK)
Ying Liu (USA)
Elena Marchiori (The Netherlands)
Andrew Martin (UK)
Jason Moore (USA)
Pablo Moscato (Australia)
Jagath Rajapakse (Singapore)
Menaka Rajapakse (Singapore)
Michael Raymer (USA)
Vic J. Rayward-Smith (UK)
Jem Rowland (UK)
Marylyn Ritchie (USA)
Ugur Sezerman (Turkey)
El-Ghazali Talbi (France)
Andrea Tettamanzi (Italy)
Janet Wiles (Australia)
Andreas Zell (Germany)
Eckart Zitzler (Switzerland)

Table of Contents

Identifying Regulatory Sites Using Neighborhood Species

Claudia Angelini[1], Luisa Cutillo[1], Italia De Feis[1], Richard van der Wath[2], and Pietro Lio'[2,*]

[1] Istituto per le Applicazioni del Calcolo "Mauro Picone" CNR, Napoly Italy
c.angelini@iac.cnr.it, cutillo@na.iac.cnr.it, i.defeis@iac.cnr.it
[2] Computer Laboratory, University of Cambridge, Cambridge UK
rcv23@cam.ac.uk, pl219@cam.ac.uk

Abstract. The annotation of transcription binding sites in new sequenced genomes is an important and challenging problem. We have previously shown how a regression model that linearly relates gene expression levels to the matching scores of nucleotide patterns allows us to identify DNA-binding sites from a collection of co-regulated genes and their nearby non-coding DNA sequences. Our methodology uses Bayesian models and stochastic search techniques to select transcription factor binding site candidates. Here we show that this methodology allows us to identify binding sites in nearby species. We present examples of annotation crossing from *Schizosaccharomyces pombe* to *Schizosaccharomyces japonicus*. We found that the eng1 motif is also regulating a set of 9 genes in *S. japonicus*. Our framework may have an effective interest in conveying information in the annotation process of a new species. Finally we discuss a number of statistical and biological issues related to the identification of binding sites through covariates of genes expression and sequences.

1 Introduction

The identification of the repertoire of regulatory elements in a genome is one of the major challenges in modern biology. Gene transcription is determined by the interaction between transcription factors and their binding sites, called motifs or cis-regulatory elements. In eukaryotes the regulation of gene expression is highly complex and often occurs through the coordinated action of multiple transcription factors. This combinatorial regulation has several advantages; it controls gene expression in response to a variety of signals from the environment and allows the use of a limited number of transcription factors to create many combinations of regulators. Identification of the regulatory elements is necessary for understanding mechanisms of cellular processes. In eukaryotes these sites comprise short DNA stretches often found within non-coding upstream regions. DNA microarrays provide a simple and natural vehicle for exploring the regulation of thousands of genes and their interactions. Genes with similar expression

* Corresponding author.

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 1–10, 2007.

profiles are likely to have similar regulatory mechanisms. A close inspection of their promoter sequences may therefore reveal nucleotide patterns that are relevant to their regulation.

In order to identify regulative sites several authors have used the following strategy: 1) candidate motifs can be obtained from the upstream regions of the most induced or most repressed genes; 2) a score can be assigned to reflect the matching of each motif to a particular upstream sequence; 3) regression analysis and variable selection methods can be used to detect sets of motifs acting together to affect the expression of genes [4,5,8].

Most of the current focus on microarray analysis is on integrating results from repeated experiments using the same species or using different species. This paper is extending this focus to transcription factor binding site identification. Following [8], we propose the use of Bayesian variable selection models to use the gene expression of an organism to find transcription binding sites of a closely related species or of a different strain. Variable selection methods use a latent binary vector to index all possible sets of variables (patterns). Stochastic search techniques are then used to explore the high-dimensional variable space and identify sets that best predict the response variable (expression). The method provides joint posterior probabilities of sets of patterns, as well as marginal posterior probabilities for the inclusion of single nucleotide patterns. We have chosen to exemplify our methodology using *S. japonicus* and *S. pombe* genomes and microarray data from cell cycle-regulated gene experiments [6].

Similar to a better known Schizosaccharomyces *S. pombe*, which has been a major model organism for cell cycle and cell biology research for thirty years, *S. japonicus* is a simple, unicellular yeast. Unlike the cousin, it readily adopts a invasive, hyphal growth form. Such growth is an important virulence trait in pathogenic fungi, making *S. japonicus* a potentially important model for fungal disease. The comparison of the *S. pombe* genome, which was sequenced several years ago, with those of its close relatives will greatly improve our understanding of the genomes and the proteins they encode. In addition, the three fission yeasts form an early-branching clade among the Ascomycete (ascus-forming) fungi, which includes yeast, hyphal fungi, and truffles [2]. Although a great deal of molecular information is available from *S. pombe*, a model eukaryote, very little is known about the *S. japonicus* cell-cycle regulative network.

Here we show that our methodology allows us to identify binding sites in *S. japonicus* using *S. pombe* gene expression data. As an example of annotation crossing from *S. pombe* to *S. japonicus* we focus on the Eng1 cluster, a set of very strongly cell cycle-regulated genes, which in *S. pombe* contains nine genes, involved in cell separation [6]. The genes are adg1 and adg2 (cell surface glycoproteins), adg3 (glucosidase), agn1 and eng1 (glycosyl hydrolases), cfh4 (chitin synthase regulatory factor), mid2 (an anillin needed for cell division and septin organization), ace2 (a cell cycle transcription factor), and SPCC306.11, a sequence orphan of unknown function. Motif searches showed that each gene of the cluster has at least one binding site for the Ace2 transcription factor (consensus CCAGCC). The Eng1 cluster has a recognizably similar functional cluster in

S. cerevisiae, the SIC1 cluster, which also contains the glycosyl hydrolase eng1 in *S. pombe*, and its ortholog DSE4 in *S. cerevisiae*.

In the next section we briefly describe the data and provide details on the statistical procedures used. Then we describe the analysis and related findings. Finally we discuss statistical issues related to the procedure we have used and the potentialities, which are currently addressed.

2 Methodology

2.1 Motif Selection Procedure

We propose a method for finding DNA binding sites which is an extension of that from [8]. While these authors have shown that variable selection is more effective than the linear regression used by [4], we have extended their procedure to time series analysis and to the use of gene expression from different species/strains. We briefly describe our methodology, pointing to the differences with respect to [8].We consider microarray experiments that explore the transcriptional responses of the fission yeast *S. pombe* [6] to cell cycle. This allows us to compare two organisms with similar biological complexity. We focus on two stress conditions in wild type cultures: oxidative stress caused by hydrogen peroxide and heat shock caused by temperature increase. Our other data consists of the organisms' genome sequences. *S. Pombe* DNA sequence data were obtained from the NCBI's FTP site (`ftp://ftp.ncbi.nih.gov/genomes/`); *S. japonicus* sequence data from (`http://www.broad.mit.edu`). The motif finding algorithms are sensitive to noise, which increases with the size of upstream sequences examined. As reported by [9], the vast majority of the yeast regulatory sites from the TRANSFAC database are located within 800 bp from the translation start site. We therefore extracted sequences up to 800 bp upstream, shortening them, if necessary, to avoid any overlap with adjacent ORF's. For genes with negative orientation, this was done taking the reverse complement of the sequences. Then we used MDScan [5] to search for nucleotide patterns. The algorithm starts by enumerating each segment of width w (seed) in the top t sequences. For each seed, it looks for w-mers with at least n base pair matches in the t sequences. These are used to form a motif matrix and the highest scoring seeds are retained. The updating step is done iteratively by scanning all w-mers in the remaining sequences and adding in or removing from the weight matrix segments that increase the score. This is repeated until the alignment stabilizes. The score of each motif is computed as in [4]. For each organism, the entire genome regions were extracted and used as background models. We searched for nucleotide patterns of length 5 to 15 bp and considered up to 30 distinct candidates for each width.

Our methodology is summarized in Figure 1 and proceed as follows: we first select candidate motifs which are generated from the over-represented nucleotide patterns, then we derive pattern scores for each motif following [4]. We continue by fitting a linear regression model relating gene expression levels (Y) to pattern scores (X), and using a Bayesian variable selection method we select motifs that best explain and predict the changes in expression level.

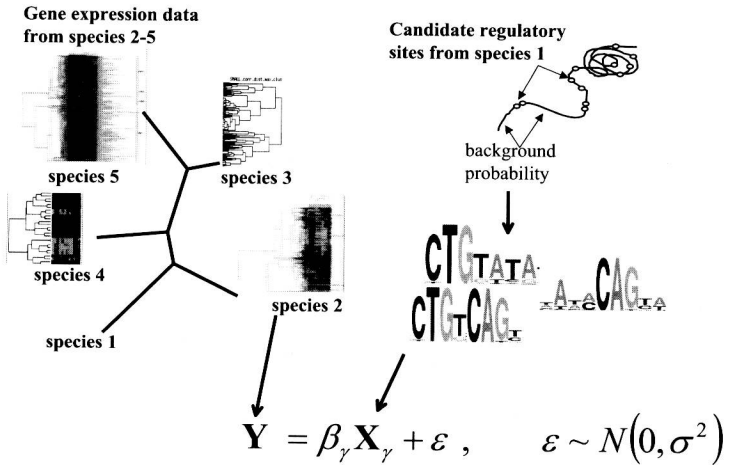

Fig. 1. Graphical representation of methodology

The variable selection method proceeds as follows. A latent vector, $\boldsymbol{\gamma}$, with binary entries is introduced to identify variables included in the model; γ_j takes on value 1 if the j^{th} variable (motif) is included and 0 otherwise. The regression model is then given by:

$$Y = X_\gamma \beta_\gamma + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2), \tag{1}$$

where the columns of X and Y are mean-centered and (γ) indexes variables included in the model [1].

We specify Bernoulli priors for the elements of $\boldsymbol{\gamma}$:

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{p} \theta^{\gamma_j} (1-\theta)^{1-\gamma_j}, \tag{2}$$

where $\theta = p_{\text{prior}}/p$ and p_{prior} is the number of covariates expected *a priori* to be included in the model. For the other model parameters, we take

$$\begin{aligned} \beta_\gamma &\sim N(0, c\{X'_\gamma X_\gamma\}^{-1}) \\ \sigma^2 &\sim \text{Inv}{-}\chi^2(a, b), \end{aligned} \tag{3}$$

where a, b and c need to be assessed through a sensitivity analysis (see [8]). The scaling value b was taken to be comparable in size to the expected error variance of the standardized data.

Note that, with respect to [8], the choice of the prior is not completely random: it draws suggestions from the motifs already discovered in analysis of the genome data of close species.

2.2 Stochastic Search

Having set the prior distributions, a Bayesian analysis proceeds by updating the prior beliefs with information that comes from the data. Our interest is in the posterior distribution of the vector $\boldsymbol{\gamma}$ given the data, $f(\boldsymbol{\gamma}|X,Y)$. Vector values with high probability identify the most promising sets of candidate motifs. Given the large number of possible vector values (2^p possibilities with p covariates), we use a stochastic search Markov chain Monte Carlo (MCMC) technique to search for sets with high posterior probabilities.

Our method visits a sequence of models that differ successively in one or two variables. At each iteration, a candidate model, $\boldsymbol{\gamma}^{\text{new}}$, is generated by randomly choosing one of these two transition moves:

(i) Add or delete one variable from $\boldsymbol{\gamma}^{\text{old}}$.

(ii) Swap the inclusion status of two variables in $\boldsymbol{\gamma}^{\text{old}}$.

The proposed $\boldsymbol{\gamma}^{\text{new}}$ is accepted with a probability that depends on the ratio of the relative posterior probabilities of the new versus the previously visited models:

$$\min\left\{\frac{f(\boldsymbol{\gamma}^{\text{new}}|X,Y)}{f(\boldsymbol{\gamma}^{\text{old}}|X,Y)},1\right\}, \tag{4}$$

which leads to the retention of the more probable set of patterns [8,1].

Our stochastic search results in a list of visited sets and corresponding relative posterior probabilities. The marginal posterior probability of inclusion for a single motif j, $P(\gamma_j = 1|X,Y)$, can be computed from the posterior probabilities of the visited models:

$$\begin{aligned} p(\gamma_j = 1|X,Y) &= \int p\left(\gamma_j = 1, \boldsymbol{\gamma}_{(-j)}|X,Y\right) d\boldsymbol{\gamma}_{(-j)} \\ &\propto \int p\left(Y|X,\gamma_j = 1,\boldsymbol{\gamma}_{(-j)}\right)\cdot p(\boldsymbol{\gamma})d\boldsymbol{\gamma}_{(-j)} \\ &\approx \sum_{t=1}^{M} p\left(Y|X,\gamma_j = 1,\boldsymbol{\gamma}_{(-j)}^{(t)}\right)\cdot p\left(\gamma_j = 1,\boldsymbol{\gamma}_{(-j)}^{(t)}\right), \end{aligned} \tag{5}$$

where $\boldsymbol{\gamma}_{(-j)}^{(t)}$ is the vector $\boldsymbol{\gamma}$ at the t^{th} iteration without the j^{th} motif.

For each organism and stress condition, we regressed the expression levels on the pattern scores using separate models. In all cases, the analyses were started with a set of around 200 patterns. We chose $p_{\text{prior}} = 10$ for the prior of $\boldsymbol{\gamma}$. This means that we expect models with relatively few motifs to perform well.

For every regression model, we ran 8 parallel MCMC chains. The searches were started with a randomly selected γ_j's set to one. We pooled together the sets of patterns visited by the 8 MCMC chains and computed the normalized posterior probabilities of each distinct visited set. We also computed the marginal posterior probabilities, $P(\gamma_j = 1|X,Y)$, for the inclusion of single nucleotide patterns.

For comparison, we repeated the analysis with MotifRegressor [4], which uses stepwise regression to select motifs.