

DISS. ETH NO. 17386

**SEARCH HEURISTICS FOR MODULE
IDENTIFICATION FROM BIOLOGICAL
HIGH-THROUGHPUT DATA**

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
STEFAN BLEULER
Dipl. El.-Ing., ETH Zurich
born July 13, 1977
citizen of
Zollikon, ZH

accepted on the recommendation of
Prof. Dr. Eckart Zitzler, examiner
Prof. Dr. Peter Bühlmann, co-examiner

2007



Institut für Technische Informatik und Kommunikationsnetze
Computer Engineering and Networks Laboratory

TIK-SCHRIFTENREIHE NR. 91

Stefan Bleuler

Search Heuristics for Module Identification from Biological High-Throughput Data



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

A dissertation submitted to
ETH Zurich
for the degree of Doctor of Sciences

Diss. ETH No. 17386

Prof. Dr. Eckart Zitzler, examiner
Prof. Dr. Peter Bühlmann, co-examiner

Examination date: August 21, 2007

Berichte aus der Informatik

Stefan Bleuler

**Search Heuristics for Module Identification
from Biological High-Throughput Data**

Shaker Verlag
Aachen 2008

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Zugl.: Zürich, ETH, Diss., 2007

Copyright Shaker Verlag 2008

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8322-6928-9

ISSN 0945-0807

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen

Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9

Internet: www.shaker.de • e-mail: info@shaker.de

Acknowledgements

Today, scientific results are rarely achieved by one person alone, much less so in an inter-disciplinary field like the one I was working in. In contrary, much of my research has been performed in close collaboration with people from different groups. Inevitably, various colleagues have made contributions to the results reported in this thesis; they are gratefully acknowledged.

- Chapter 3 and Appendix B: Amela Prelić and Eckart Zitzler developed and implemented the reference method Bimax and performed the running-time analysis. Amela Prelić performed the experiments for the validation on the real data and a part of the experiments on the synthetic data sets.
- Chapter 5: Markus Friberg performed the search for the transcription factor binding sites and Philip Zimmermann provided the biological interpretation of the biclustering results.
- Chapter 6: In the course of his masters thesis which I supervised, Michael Calonder implemented a part of the algorithm for multi-objective biclustering and performed some of the experiments reported.
- Appendix A: Lothar Thiele provided the experimental results.

Besides these specific contributions, many others have contributed in various ways, by providing data and tools or through helpful discussions and advice. In particular, I would like to thank

- my advisor Eckart Zitzler for his guidance on my scientific journey during which he always provided me with just the right combination of support and independence,
- Peter Bühlmann for kindly serving as co-examiner and for his many valuable inputs during our collaboration in the REP project,
- my colleagues from the Reverse Engineering Project for this great experience in interdisciplinary research,
- Nicola Zamboni for the good collaboration on the fluxome project, and
- my colleagues at the Computer Engineering Laboratory for the great environment they provided both personally and scientifically.

Contents

Contents	vii
1 Introduction	1
1.1 Biological Motivation	1
1.2 Research Questions	2
1.3 Contributions	4
1.4 Overview	5
2 Background and Related Work	7
2.1 High-Throughput Measurements in Cell Biology	7
2.1.1 Genome	8
2.1.2 Transcriptome	8
2.1.3 Proteome	9
2.1.4 Metabolome	10
2.1.5 Annotations	11
2.2 A General Problem Formulation of Module Identification	11
2.3 Methods for Module Identification	12
2.3.1 A Categorization of Approaches	13
2.3.2 Overview of Existing Biclustering Methods	14
2.3.3 Query Gene Methods	17
2.4 Randomized Search Algorithms	18
3 Empirical Validation of the Biclustering Concept	21
3.1 Motivation	21
3.2 Related Work	22
3.3 Biclustering Methods	24
3.3.1 Selected Algorithms	24
3.3.2 Reference Method (Bimax)	24
3.4 Comparison Methodology	27
3.4.1 Validation Using Synthetic Data	28
3.4.2 Validation Using Prior Knowledge	30
3.4.3 Implementation Issues	32
3.5 Results	33

3.5.1	Synthetic Data	33
3.5.2	Real Data	42
3.6	Summary	44
4	An Evolutionary Algorithm Framework for Biclustering	47
4.1	Motivation	47
4.2	Related Work	48
4.3	Model	49
4.4	Evolutionary Algorithm	50
4.5	Local Search Algorithm	54
4.6	Simulation Results	55
4.6.1	Data Sets and Experimental Setup	55
4.6.2	Finding One Bicluster	56
4.6.3	Finding a Set of Biclusters	61
4.7	Summary	65
5	Biclustering of Multiple Gene Expression Data Sets	67
5.1	Motivation	67
5.2	Model	69
5.2.1	A Homogeneity Score for Trends	69
5.2.2	A Biclustering Model for Multiple Data Sets	70
5.3	Optimization Algorithm	72
5.4	Experimental Results	73
5.4.1	Experimental Setup	74
5.4.2	Comparison to Alternative Algorithms	77
5.4.3	Effects of Combining Data Sets	83
5.4.4	Differential Coexpression	86
5.4.5	Biological Content of Exemplary Biclusters	88
5.5	Summary	91
6	Biclustering of Multiple Types of Biological High-Throughput Data	93
6.1	Motivation	93
6.2	Related Work	95
6.2.1	Data Integration	95
6.2.2	Multiobjective Clustering	95
6.3	Model	96
6.4	Optimization Algorithm	98
6.5	Results	99
6.5.1	Experimental Setup	99
6.5.2	Performance of the Proposed Algorithm	100
6.5.3	Application to Different Biological Scenarios	103
6.6	Summary	106

7 Identification of Characteristic Differences in Fluxome Measurements	109
7.1 Motivation	109
7.2 Related Work	111
7.3 Model	111
7.4 Optimization Algorithm	114
7.5 Results	116
7.5.1 Data Preparation and Experimental Setup	117
7.5.2 Evaluation of the Evolutionary Algorithm	117
7.5.3 Validation using Flux Ratios	119
7.6 Summary	119
8 Conclusions	125
8.1 Key Results	125
8.2 Outlook	127
A A Portable Interface for Search Algorithms	131
A.1 Motivation	131
A.2 Design Goals and Requirements	134
A.3 Architecture	135
A.3.1 Control Flow	135
A.3.2 Data Flow	137
A.4 Implementation Aspects	138
A.4.1 Synchronization	138
A.4.2 Data Exchange	139
A.4.3 Parameters	141
A.4.4 Correctness	141
A.5 Experimental Results	143
A.6 Summary	144
B Running-Time Analyses for Chapter 3	149
B.1 Bimax (Reference Method)	149
B.1.1 Algorithm	149
B.1.2 Running-Time Analysis	151
B.2 Incremental Procedure	152
B.2.1 Algorithm	152
B.2.2 Running-Time Analysis	153
C List of Acronyms	155
D List of Symbols	157
Bibliography	161

Curriculum Vitae

177