Algorithmic Data Analysis (Algodan) Centre-of-Excellence

Triennal report 2011-2013

April 16, 2014

Esko Ukkonen, Teija Kujala, Pirjo Moen, eds.

Preface

The Finnish Centre of Excellence in Algorithmic Data Analysis Research (Algodan) has completed its six year term. This report describes the activities of the Centre in 2011-13, and also looks forward into the future of the research groups after Algodan. Moreover, the report presents an extensive bibliometric analysis of the publications produced by the Centre during its total term 2008-13. Research activities are reported according to the teams and their research groups. The groups present their members, mission and main results, cooperation and societal impact as well as a list of most important publications. Complete lists of publications and PhD degrees are given at the end of the report.

I would like to thank all members of the Centre for their efforts over the years to make our Algodan a success.

Helsinki, April 16, 2014

Esko Ukkonen

www.cs.helsinki.fi/research/algodan/

Contents

1. Introduction

Summary of Algodan centre as described in the original application

The importance of data analysis in science and in industry is increasing continuously, as our ability to measure and store data grows. While data analysis is as old as science itself, the new methods of collecting raw data pose unprecedented challenges and opportunities to data analysis and to the algorithms of data analysis.

The Algorithmic Data Analysis (Algodan) Centre of Excellence develops new concepts, algorithms, principles, and frameworks for data analysis. The work combines strong basic research in computer science with interdisciplinary work in a wide variety of scientific disciplines and industrial problems.

The research of the Algodan CoE lies in the areas of combinatorial pattern matching, data mining, and machine learning. The work in Algodan is strongly interdisciplinary: we cooperate constantly with application experts in various application areas, formulating novel computational concepts and ways of attacking the scientific and industrial problems of the application areas. Developing new concepts and algorithms is an iterative process consisting of interacting extensively with the application experts, formulating computational concepts, analyzing the properties of the concepts, designing algorithms and analyzing their performance, implementing and experimenting with the algorithms, and applying the results in practice. The main application areas of the Algodan CoE are in biology, medicine, telecommunications, environmental studies, linguistics, and neuroscience.

The formulation of new computational concepts, their analysis, and the design of algorithms are some key ingredients that make the Algodan CoE unique. First, rather than concentrating on improvements to existing problems and methods, the CoE focuses on defining new tasks where significant impact can be made by introducing new concepts. Second, we emphasize the need for analyzing the performance of the algorithms, instead of just relying on heuristic approaches. Third, we use our strong background in algorithmic and probabilistic methods to guarantee that our algorithms perform well both in terms of modelling accuracy and robustness, and in terms of computational complexity and practical efficiency.

The research in Algodan is grouped under four interacting themes: sequence analysis, learning from and mining complex and heterogeneous data, discovery of hidden structure in high-dimensional data, and foundations of algorithmic data analysis. All these themes combine aspects of combinatorial pattern matching, data mining, and machine learning.

The host organizations of the Algodan CoE are University of Helsinki and Aalto University[1]. The CoE is in part a continuation of the "From Data to Knowledge" CoE (2002-2007), and consists of about 70 persons. The director of the Algodan CoE is Professor Esko Ukkonen and the vice-director is Vice President, Professor Heikki Mannila[2].

[1] Until 31st of December 2009 Helsinki University of Technology.
[2] From 1st of March 2012 Professor Heikki Mannila was appointed as the President of the Academy of Finland.

Main research themes

The main research themes of the Algodan CoE are the following.

- S – Sequence analysis
- L – Learning from and mining structured and heterogeneous data
- D – Discovery of hidden structure in high-dimensional data
- F – Foundations of algorithmic data analysis

There is considerable overlap between the themes: certain algorithmic and probabilistic techniques occur in many themes. In the same way, several themes can be used for a single application. We next describe the themes briefly.

Sequence analysis considers the algorithmic techniques for sequential data. The key methods in the theme are string algorithms, pattern discovery techniques, dynamic programming, and probabilistic modelling. Examples of the algorithmic tasks in the area are approximate string matching, episode discovery, and finding motifs and orders from data. The techniques of sequence analysis have numerous applications in, for example, gene mapping, finding regulatory regions in genomes, telecommunications, linguistics, and paleontology.

Most applications have multiple types of data objects, many different types of data, etc., instead of the classical situation of a single table with observations and variables. Learning from and mining structured and heterogeneous data looks for techniques for data analysis tasks involving such data sets. The methods studied are pattern discovery, prediction of structured objects, the analysis of flows, etc. The applications include biological data analysis, information retrieval, telecommunications, and environmental studies. Algorithmic techniques for probabilistic modelling are crucial in this theme.

The high dimensionality of many datasets causes interesting modelling problems and leads to extremely challenging algorithmic questions. The third theme, discovery of hidden structure in high-dimensional data, looks at how to find latent structure in high-dimensional data sets. The latent structure can be in the form of components, as in independent component analysis, or cluster-like structures, or it can be a parsimonious model giving weight only to a small fraction of the observed variables. The techniques in this theme are based on probabilistic modelling, with a strong algorithmic component.

The theme on foundations of algorithmic data analysis looks at the frameworks of algorithmic data analysis. What can be said about the limitations of pattern discovery? What are the fundamental bounds on the efficiency of string algorithms? What is the computational complexity of fitting probabilistic models of a certain type? Questions such as these abound in algorithmic data analysis, and they are fascinating problems in core computer science.

2. Analysis of ALGODAN Publications 2008-2013

The total number of citations to Algodan publications in 2008-2013 found in Scopus is 2673 and in Publish or Perish (Google Scholar) 6166. The analysis was made by Helsinki University Library bibliometrics team in April 2014 using the publication list provided by the research group. For full report see Appendix 1.

Summary of publication statistics

There are 661 publications listed for 2008-2014. The types and annual statistics are summarized in Table 1.

Table 1: ALGODAN publications by year and type. See explanation below for type classification.

Year	A1	A3-A4	B-E	C	F	G	Total
2008	40	68	10	0	0	6	124
2009	34	46	9	1	0	7	97
2010	34	75	15	0	0	5	129
2011	31	52	6	0	0	3	92
2012	29	51	20	0	1	8	109
2013	32	54	4	0	10	8	108
2014						2	2
Total	200	346	64	1	11	39	661

The publications are classified according to the following scheme:

A1	Articles in refereed scientific journals
A3-A4	Refereed conference articles and articles in edited books
B-E	Technical reports and other publications
C	Books
F	Artistic works
G	Theses

Citation analysis with Scopus

The ALGODAN publications found in Scopus received 2673 citations by beginning of April 2014. A more detailed view for each year is seen in Table 2.

Table 2: Citations in Scopus by year

Year	2008	2009	2010	2011	2012	2013	2014	Total
2007	1	0	0	0	0	0	0	1
2008	22	96	162	158	180	133	26	777
2009		37	152	167	159	146	27	688
2010		2	53	243	263	266	39	866
2011		1	0	8	69	92	13	183
2012					26	80	14	120
2013						30	8	38
2014							0	0
Total	23	136	367	576	697	747	127	2673

Citation analysis with Publish or Perish

Citations from Publish or Perish (based on Google Scholar) are listed in Table 3. The quality of the citations has not been controlled in any way.

Table 3: Citations by year and type according to PoP

Year	A1	A3-A4	B-E	C	Total
2008	1111	752	23	0	1886
2009	1006	532	9	208	1755
2010	1102	386	22	0	1510
2011	270	166	0	0	436
2012	175	136	25	0	336
2013	149	65	29	0	243
Total	3813	2037	108	208	6166

3. Funding of ALGODAN in 2008-2013

Table 4: Funding spent by the unit by year and source

Funding agency	2008	2009	2010	2011	2012	2013
Academy of Finland	1 548 000	1 319 000	1 518 000	1 604 000	1 178 000	1 304 000
Tekes	46 000	223 000	215 000	5 000	108 000	65 000
EU	362 000	340 000	59 000	69 000	113 000	235 000
Other	79 000	212 000	241 000	13 000	3 000	0
Own funding	1 055 000	1 324 000	1 452 000	1 103 000	1 197 000	810 000
Ministry	2 000	66 000	0	0	25 000	50 000
TOTAL	3 092 000	3 484 000	3 485 000	2 794 000	2 624 000	2 464 000

4. Reports from the Teams and their Groups

Team Data Mining: Theory and Applications

Data mining – theory and applications

Members

- Kai Puolamäki, PhD, Docent, Group leader (-2012)
- Aristides Gionis, Associate Professor, Group leader (2013-)
- Heikki Mannila, Professor (on leave of absence)
- Panagiotis Papapetrou, PhD, Postdoctoral researcher (-2012)
- Nikolaj Tatti, PhD, Postdoctoral researcher (2013-)
- Michael Mathioudakis, PhD, Postdoctoral researcher (2013-)
- Aleksi Kallio, Part-time doctoral student
- The following people finished their PhD between 2011–2013:
    - Esa Junttila, August 2011
    - Markus Ojala, November 2011
    - Niko Vuokko, February 2012
    - Sami Hanhijärvi, May 2012
    - Jefrey Lijffijt, December 2013

Mission of the group

The Data Mining: Theory and Applications group at Aalto University conducts research on finding local patterns and global models in discrete high-dimensional data. Techniques for this task include both algorithmics in the traditional computer science sense and probabilistic methods. The group was founded by Professor Heikki Mannila who was later appointed the Vice-President of Aalto and then President of the Academy of Finland, and who still contributes to doctoral student supervision in the group. In 2013, Aristides Gionis, new faculty member in Aalto, joined the activities of the group and broadened the research agenda to new areas such as graph mining, social-network analysis, analysis of information networks.

Research activities

S - Sequence analysis

The group has been very active in analysis of sequential data. Research highlights include novel approaches to the problem of discovering episodes in sequential data [2] (continuing on a topic that was introduced by the seminal work of professor Mannila in the 90's), segmentation algorithms that scale to very large data for a wide family of models [4], methods for discovering surprising sequential patterns [5], and methods for selecting the values of parameters used in sequential pattern mining [7].

In December 2013 Jefrey Lijffijt defended his PhD thesis titled "Computational methods for comparison and exploration of event sequences". The dissertation obtained an honorary grade and it received the award for the "Best doctoral dissertation of 2013" in the Aalto University School of Science.

F - Foundations of algorithmic data analysis

The group is internationally known for contributions in pattern discovery. In the recent years the group continued successfully the research on this area. Highlights include work on sampling methods for finding robust itemsets [1] and methods to evaluate the statistical significance of pattern sets [3].

A - Applications

As the name of the group suggests, there is strong emphasis on applications and multidisciplinary research. This is demonstrated with the active collaboration of the group members with scientists from different domains (biology, paleontology, linguistics), and cross-disciplinary publications [6,8,9].

Future plans

The group is internationally very well known in the field of data mining and knowledge discovery. Additionally, the group has had a substantial impact on application areas – ecology and linguistics being the recent focuses – where the group has introduced new computational concepts and developed methods together with the application area experts.

Prof Heikki Mannila, who established the group, was nominated in 2009 as the Vice President of then newly established Aalto University. Since 2012, Prof Mannila has been serving as the President of the Academy of Finland.

In 2013 Aristides Gionis joined Aalto University as an associate professor, assumed the responsibility for the group, broadened the research agenda, and recruited new students and postdocs. The emphasis of the group in the next years will be in establishing its international reputation, building its collaboration network, recruiting top-notch postdoctoral researchers, and guiding students in research.

Societal, economic, and technical impact

The main societal impact of the work is in the use of the methods in other sciences. This is demonstrated with successful cross-disciplinary publications in the domains of biology, ecology, and linguistics.

There is a high degree of mobility and high rate of renewal. Aristides Gionis joined the group after being a senior research scientist in Yahoo! Research. Nikolaj Tatti returned to Aalto after a successful postdoc position in Belgium, and Michael Mathioudakis obtained his PhD in 2013 from University of Toronto.

The alumni of the group have positioned themselves well in industry and academia. Prof Mannila is the President of the Academy of Finland. Kai Puolamäki is a research scientist at the Finnish Institute of Occupational Health, and Panagiotis Papapetrou obtained an associate professor position in Stockholm University. Recent PhD graduates, Niko Vuokko, Sami Hanhijärvi, Esa Junttila, and Markus Ojala have all obtained attractive positions in the pioneering Helsinki data-science sector.

Cooperation

The group works heavily with other teams of the Algodan center, and there is also a wide international cooperative network, as evidenced by the publication list.

Selected publications

1. Nikolaj Tatti, Fabian Moerchen, Toon Calders. Finding Robust Itemsets under Subsampling. ACM Transactions on Database Systems, to appear.
2. Nikolaj Tatti. Discovering Episodes with Compact Minimal Windows. Data Mining and Knowledge Discovery, *to appear*.
3. Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki. A statistical significance testing approach to mining the most informative set of patterns. Data Mining and Knowledge Discovery, 28(1):238–263, 2014.
4. Nikolaj Tatti. Fast sequence segmentation using log-linear models. Data Mining and Knowledge Discovery, 27(3):421–441, 2013.

5. Jefrey Lijffijt. A fast and simple method for mining subsequences with surprising event counts. In ECML-PKDD, pp. 385–400, 2013.
6. Hannes Gamper, Christina Dicke, Mark Billinghurst, and Kai Puolamäki. Sound sample detection and numerosity estimation using auditory display. ACM Transactions on Applied Perception, 10(1):4, 2013.
7. Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki. Size matters: Finding the most informative set of window lengths. In ECML-PKDD, pp 451–466, 2012.
8. Liping Liu, Kai Puolamäki, Jussi T. Eronen, Majid M. Ataabadi, Elina Hernesniemi, and Mikael Fortelius. Dental functional traits of mammals resolve productivity in terrestrial ecosystems past and present. Proceedings of the Royal Society B, 279(1739):2793–2799, 2012.
9. Panagiotis Papapetrou, Gary Benson, and George Kollios. Mining Poly-regions in DNA. International Journal of Data Mining and Bioinformatics, 6(4):406–428, 2012.
10. Alexios Kotsifakos, Panagiotis Papapetrou, Jaakko Hollmén, Dimitrios Gunopulos, Vassilis Athitsos, and George Kollios. Hum-a-song: A Subsequence Matching with Gaps-Range-Tolerances Query-By-Humming System. Proceedings of the VLDB Endowement, 4(11):1930–1933, 2012.

Parsimonious Modelling Group

Members

- Jaakko Hollmén, Chief Research Scientist, Group leader
- Indrė Žliobaitė, PhD, Postdoctoral researcher (2013-)
- Jesse Read, PhD, Postdoctoral researcher (2013-)
- Mika Sulkava, Academy Postdoctoral Researcher (-2011)
- Miguel Angel Prada, PhD (-2010)
- Mikko Korpela, MSc (Tech.), Doctoral student
- Janne Toivola, MSc (Tech.), Doctoral student (-2013)
- Prem Raj Adhikari, MSc (Tech.), Doctoral student
- Olli-Pekka Rinta-Koski, M.Sc. (Tech.), Doctoral student (-2013)

Mission of the group

The research group Parsimonious Modelling develops novel computational data analysis methods and applies these methods on two application fields: cancer genomics and environmental informatics. Parsimonious modeling aims at simple, compact, or sparse models as a result of learning from data in the presence of very little or no a priori information about the modeled problem. Simplicity of the models facilitates understanding of the problem domain by humans.

Research activities

In the area of *cancer genomics*, the research concentrated on the analysis of high throughput microarray data, such as gene expression data and array-based chromosomal genomic hybridization (aCGH) data [13, 16]. A clear emphasis is on the aCGH data measuring gene-specific genomic aberrations, whereas gene expression data has been employed when integrating data sets together in joint analysis scenario. Multi-resolution data has been an active topic of research [1, 14, 15, 19]. Health and wellbeing applications have been surveyed [20].

Methodologically, the research concentrates on biomarker selection problems [13], model selection criteria in search-based feature selection, as well as modeling of multiresolution data [1, 14, 15, 19].

The second research area of the group, *environmental informatics* is understood as the analysis of time series from the natural environment (such as forests, trees, and climate) as well as the man-made, built environment. The analysis of the data from the built environment is strengthened by the acquisition of a research project *TrafficSense* - Energy efficient traffic with crowdsensing.

Projects on the natural environment focused on the forests and their role in the carbon balance [8], environmental monitoring [10], understanding factors behind tree growth [5] and the analysis of proxy time series for climate reconstructions. The man-made environment currently embodies structures, such as buildings and bridges, for instance, which can be equipped with measurement sensors to yield large data bases reflecting health of the structures. The analysis is concerned with identifying or discovering abstract states for the structure and the problem is to detect abnormal states and diagnose faults [9, 11, 12].

Generic methodological research in time-series and sequence analysis has been conducted with other Algodan researchers [6] and with others [7, 17]. The research group has been actively involved in conference organization activities [3, 4].

Publication activity in the group has been very good. The group has hosted several visits between 2011 and 2013. Good balance between applications and methodologies has been achieved.

Future plans

Research will be continued in all areas, with more emphasis on the environmental applications. The recent project acquisition of TrafficSense- Energy efficient traffic with crowdsensing will undoubtedly guide the research into transportation related themes. Multiresolution data analysis is now investigated in both cancer genomics and environmental informatics, which could bring synergetic effects.

Societal, economic, and technical impact

The research group has taken an active role in conference organization of the IDA and DS conference series for many years. The application oriented research papers contrinute to the fields of environmental sciences and transportation research.

Cooperation

- Sakari Knuutila, Laboratory of Cytomolecular Genetics (CMG), University of Helsinki, Finland: Joint projects on cancer genomics
- Harri Mäkinen and Pekka Nöjd, Finnish Forest Research Institute, Vantaa, Finland: Joint research on forest growth and proxy time series
- Pertti Hari and Eero Nikinmaa, University of Helsinki, Department of Forestry, Helsinki, Finland: Joint research on forest growth and proxy time series
- Sebastiaan Luyssaert and Ivan Janssens, University of Antwerp, Belgium: Joint research on carbon balance and the role of forests
- Dimitrios Gunopulos, University of Athens, Greece: Joint research on sequence analysis, string matching and analysis of the built environment

Selected publications

1. Prem Raj Adhikari and Jaakko Hollmén. Patterns from Multi-Resolution 0-1 Data. In Bart Goethals, Nikolaj Tatti, and Jilles Vreeken, editors, In Proceedings of the ACM SIGKDD Workshop on Useful Patterns (UP'10), pages 8—12. July 25, 2010. Washington, DC, USA.
2. Serafin Alonso and Mika Sulkava and Miguel Angel Prada and Manuel Dominguez and Jaakko Hollmén. Comparative analysis of power consumption in university building using envSOM. In Advances in Intelligent Data Analysis X — Proceedings of the 10th International Symposium (IDA 2011), Volume 7014 of Lecture Notes in Computer Science. Pages 10—21, Springer-Verlag, October 2011. Porto, Portugal.
3. Tapio Elomaa, Jaakko Hollmén, and Heikki Mannila, editors. Discovery Science — Proceedings of the 14th International Conference (DS 2011), volume 6926 of Lecture Notes in Computer Science. Springer-Verlag, October 2011.
4. João Gama, Elizabeth Bradley, and Jaakko Hollmén, editors. Advances in Intelligent Data Analysis — Proceedings of the 10th International Symposium on Intelligent Data Analysis (IDA 2011), volume 7014 of Lecture Notes in Computer Science. Springer-Verlag, October 2011.
5. Mikko Korpela, Pekka Nöjd, Jaakko Hollmén, Harri Mäkinen, Mika Sulkava, and Pertti Hari. Photosynthesis, temperature and radial growth of Scots Pine in northern Finland: identifying the influential time intervals, Trees — Structure and Function. 25(2):323–332, April, 2011.
6. Orestis Kostakis, Panagiotis Papapetrou, and Jaakko Hollmén. ARTEMIS: Assessing the similarity of event-interval sequences. In Proceedings of the European Conference of Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML/PKDD), Volume 6912 of Lecture Notes in Computer Science. Pages 229—244, Springer-Verlag, September 2011.

7. Alexios Kotsifakos, Panagiotis Papapetrou, Jaakko Hollmén, and Dimitrios Gunopulos. A subsequence matching with gaps-range-tolerances framework: A query-by-humming application. In Proceedings of the Very Large Database Endowment (PVLDB), (4)11:761–771, August 2011.
8. S. Luyssaert, P. Ciais, S. L. Piao, E.-D. Schulze, M. Jung, S. Zaehle, M. J. Schelhaas, M. Reichstein, G. Churkina, D. Papale, G. Abril, C. Beer, J. Grace, D. Loustau, G. Matteucci, F. Magnani, G. J. Nabuurs, H. Verbeeck, M. Sulkava, G. R. van der Werf, and I. A. Janssens. The European carbon balance. Part 3: forests. Global Change Biology, 16(5):1429–1450, May 2010.
9. Miguel A. Prada and Janne Toivola and Jyrki Kullaa and Jaakko Hollmén. Three-way analysis of structural health monitoring data, Neurocomputing, Volume 80, Pages 119–128. March 2012.
10. Mika Sulkava, Sebastiaan Luyssaert, Sönke Zaehle, and Dario Papale. Assessing and improving the representativeness of monitoring networks: The European flux tower network example. Journal of Geophysical Research – Biogeosciences, 116:G00J04, May 2011.
11. Janne Toivola, Miguel A. Prada, and Jaakko Hollmén. Novelty detection in projected spaces for structural health monitoring. In Paul R. Cohen, Niall M. Adams, and Michael R. Berthold, editors, Advances in Intelligent Data Analysis IX, volume 6065 of LNCS, pages 208–219. Springer-Verlag. May 2010. Tucson, Arizona, USA.
12. J. Toivola and J. Hollmén. Collaborative filtering for coordinated monitoring in sensor networks. In Proceedings of the ICDMW 2011 11th IEEE International Conference on Data Mining Workshops, pages 987-994. IEEE Computer Society, December 2011. Vancouver, Canada.
13. Anu Usvasalo, Riikka Raty, Arja Harila-Saari, Pirjo Koistinen, Eeva-Riitta Savolainen, Sakari Knuutila, Erkki Elonen, Ulla M. Saarinen-Pihkala, and Jaakko Hollmén. Prognostic classification of patients with acute lymphoblastic leukemia by using gene copy number profiles identified from array-based comparative genomic hybridization data. Leukemia Research, 34(11):1476–1482, November, 2010.
14. Prem Raj Adhikari, Jaakko Hollmén, 2013. Fast Progressive Training of Mixture Models for Model Selection. Journal of Intelligent Information Systems, Springer, Published Online: 01 December 2013, In Press.
15. Prem Raj Adhikari, Jaakko Hollmén, 2012. Mixture Models from Multiresolution 0-1 Data. In J. Fürnkranz, E. Hüllermeier, and T. Higuchi, Editors, Proceedings of Sixteenth International Conference on Discovery Science (DS 2013), Volume 8140 of Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin Heidelberg, pages 1-16, October 6-9, 2013, Singapore.
16. Tarja Niini, Ilari Scheinin, Leo Lahti, Suvi Savola, Fredrik Mertens, Jaakko Hollmén, Tom Böhling, Aarne Kivioja, Karolin H. Nord, and Sakari Knuutila. Homozygous deletions of cadherin genes in chondrosarcoma — an array CGH study. Cancer Genetics, 205(11):588—593, November 2012.
17. Alexios Kotsifakos, Panagiotis Papapetrou, Jaakko Hollmén, Dimitrios Gunopulos, Vassilis Athitsos, and George Kollios. Hum-a-song: A subsequence matching with gaps-range-tolerances query-by-humming system. Proceedings of the Very Large Databases Endowement (PVLDB), 4(12):1930—1933, 2012.
18. Indrė Žliobaitė, Jaakko Hollmén. Fault tolerant regression for sensor data. In Proceedings of The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013), Volume 8188 of Lecture Notes in Computer Science, Springer-Verlag, pages 449—464, September, 2013, Prague, Czech Republic.
19. Prem Raj Adhikari, Jaakko Hollmén. Multiresolution Mixture Modeling using Merging of Mixture Components. In Proceedings of Fourth Asian Conference on Machine Learning (ACML 2012), Volume 25 of Journal of Machine Learning Research — Proceedings Track, pages 17—32, November 4-6, 2012, Singapore.
20. Olli-Pekka Rinta-Koski. Monitoring sleep quality with non-invasive sensors. Licentiate's thesis, Aalto University, March 2013.

Combinatorics, Algebra, and Computing (CO-ALCO)

Members

- Mikko Koivisto, Academy Research Fellow (8/2008-10/2013), Assistant Professor (1/2013-), Co-leader
- Petteri Kaski, Academy Research Fellow (9/2011-), Associate Professor (1/2012-), Co-leader
- Pekka Parviainen, Doctoral student (-6/2012, PhD 3/2012)
- Janne Korhonen, Doctoral student (PhD 2/2013)
- Juho-Kustaa Kangas, Doctoral student (1/2012-)
- Teppo Niinimäki, Doctoral student

The Phenomics group was merged to the CO-ALCO group in the beginning of 2013.

Mission of the group

The group develops and applies combinatorial and algebraic tools for computational problems, focusing on exact deterministic algorithms. Applications range from fundamental combinatorial problems to computational tasks associated with established probabilistic models in machine learning and data mining.

Research activities

D - discovery of hidden structure in high-dimensional data

We have continued our research on algorithmic foundations of learning graphical models from data. One one hand, we have studied parameterized problems. For special classes of graphical models called polytrees, we showed both tractability and intractability results [6]. For the class of Bayesian networks parameterized by the treewidth (of the moralized graph), we showed an intractability result but also gave a novel dynamic programming algorithm that is practical for small problem instances [8]. On the other hand, we have introduced a partial ordering based framework for learning Bayesian networks (see, e.g., the JMLR article [10] and Parviainen's PhD thesis). We have also studied more efficient ways to learn Bayesian networks using sampling-based estimators [9].

Complementary to learning graphical models, we have pursued parameterized solutions to locate connected motifs in graph data; our randomized algorithms run in time linear in the size of the host graph, and scale exponentially only in the size of the motif [3].

F - Foundations of algorithmic data analysis

We have continued our work on combinatorially flavored variants of zeta and Möbius transforms and on improved time-space tradeoffs for hard combinatorial problems. For example, we showed that every lattice with v elements, n of which are nonzero and join-irreducible, has FFT-like arithmetic circuits of size $O(vn)$ for computing the zeta transform and its inverse, thus enabling fast multiplication in the Möbius [2]. Another highlight concerns the classic subset sum problem: we gave a randomized algorithm that currently yields the best known space-time tradeoff [1]. We have also studied to what extent some recent results obtained using negation (or subtraction) can be achieved with only monotone computation, i.e., without negation: we proved both positive and negative results (see, e.g., ref. 7 and Korhonen's PhD thesis). Very recently, we show that a class of parameterized counting problems can be solved faster than “meet-in-the-middle time” by a combination of fast Möbius inversion and fast rectangular matrix multiplication [4].

Future plans

The Phenomics group was merged to the CO-ALCO group in the beginning of 2013.

Societal, economical and technical impact

As the group focuses on foundations of algorithmic data analysis, we do not expect to see high societal impact within the next five years. However, we invest substantial efforts to high-risk, high-yield research problems of relatively broad theoretical interest. We expect that some of our results will quickly prove useful for our research community and have high impact in the long run, say within the next fifty years. A specific example of visibility in the broader community is the recent review article in the Communications of the ACM [5].

Cooperation

The members of the groups are active also in research projects and study groups initiated by or shared with other groups in Algodan: combinatorial structures in binary data (the Data Mining group); genotype and phenotype analysis (the Phenomics group); causal networks, inference, and discovery (Hyvärinen, Hoyer); local algorithms (Polishchuk).

The current members of the group are all affiliated also with the Helsinki Institute for Information Technology (HIIT). There is active cooperation with other researchers, nationally and internationally:

- P. Austrin; KTH, Sweden; joint publications.
- P. Floréen (J. Suomela), HIIT; local algorithms; joint publications.
- F. Fomin, University of Bergen, Norway; algorithm theory; joint research, manuscript, visits.
- A. Hulpke, Colorado State University, USA; computational algebra; joint publications.
- T. Husfeldt (A. Björklund), Lund University, Sweden & IT University of Copenhagen, Denmark; algorithm theory; joint publications.
- M. Järvisalo; HIIT; relations to satisfiability and constraint satisfaction problems; joint publications.
- J. Nederlof, Utrecht University, the Netherlands; algorithm theory; joint publications, visits.
- S. Szeider, TU Vienna, Austria; parametrized algorithms and complexity in the context of probabilistic models; joint publications, visits.
- P. R. J. Östergård, Aalto University, Helsinki; combinatorics; joint publications.

Selected publications

1. Per Austrin, Petteri Kaski, Mikko Koivisto, and Jussi Määttä. Space-time tradeoffs for Subset Sum: An improved worst case algorithm. 40th International Colloquium on Automata, Languages, and Programming (ICALP 2013), pp. 45-56, Springer, 2013.
2. Andreas Björklund, Thore Husfeldt, Petteri Kaski, Mikko Koivisto, Jesper Nederlof, and Pekka Parviainen. Fast zeta transforms for point lattices. 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2012), pp. 1436-1444, SIAM, 2012. (Journal version to appear in ACM Trans. Alg.)
3. Andreas Björklund, Petteri Kaski, Lukasz Kowalik. Probably optimal graph motifs. 30th International Symposium on Theoretical Aspects of Computer Science (STACS 2013), pp. 20-31, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2013.
4. Andreas Björklund, Petteri Kaski, Lukasz Kowalik. Counting thin subgraphs via packings faster than meet-in-the-middle time. 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2014), pp. 594-603, SIAM, 2014.
5. Fedor Fomin and Petteri Kaski. Exact exponential algorithms. Commun. ACM 56 (2013) 80-88.

6. Serge Gaspers, Mikko Koivisto, Matthieu Liedloff, Sebastian Ordyniak, and Stefan Szeider. On finding optimal polytrees. 26th Conference on Artificial Intelligence (AAAI 2012), AAAI, 2012.
7. Matti Järvisalo, Petteri Kaski, Mikko Koivisto, and Janne Korhonen. Finding efficient circuits for ensemble computation. 15th International Conference on Theory and Applications of Satisfiability Testing (SAT 2012), pp. 369-382, Springer, 2012.
8. Janne Korhonen and Pekka Parviainen. Exact learning of bounded tree-width Bayesian networks. 16th International Conference on Artificial Intelligence and Statistics (AISTATS 2013), pp. 370-378, JMLR.org, 2013.
9. Teppo Niinimäki and Mikko Koivisto. Annealed importance sampling for structure learning in Bayesian networks. 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), IJCAI/AAAI, 2013.
10. Pekka Parviainen and Mikko Koivisto. Finding optimal Bayesian networks using precedence constraints. Journal of Machine Learning Research 14 (2013) 1387-1415.

Phenomics Group

Members

- Heikki Mannila, Professor, Group leader (-2/2012)
- Mikko Koivisto, Academy Research Fellow (8/2008-10/2013), Assistant Professor (1/2013-), Co-leader
- Pekka Parviainen, Doctoral student (-6/2012, PhD 3/2012)
- Jaana Wessman, Doctoral student (-12/2011, PhD 4/2012)
- Teppo Niinimäki, Doctoral student

The Phenomics group was merged to the CO-ALCO group in the beginning of 2013.

Mission of the group

The group develops and applies data mining techniques to identify new phenotypic and genotypic associations in population sample databases.

Research activities

D - Discovery of hidden structure in high-dimensional data

We have completed a project that aimed at detecting new, biologically more meaningful phenotypic associations using data from the Northern Finland Birth Cohort of 1966 (NFBC66); this is part of the larger Consortium for Neuropsychiatric Phenomics (coordinated by the University of California in Los Angeles). Clustering the subjects according to a set of temperament phenotypes using a mixture model method revealed four coherent clusters and interesting dependencies between the so-called temperament and character inventory subscales; the results are published in PLos ONE [1, 5]; a thorough description and discussion of the methodology, with application to other data sets are published in the PhD thesis of Jaana Wessman (2012).

We have also progressed on a related theme of analyzing causal and statistical dependencies within and between phenotypes and genotype using Bayesian network models. On one hand, in collaboration with domain experts we have prepared parts of the NFBC66 data for the analyses by pruning, merging, and discretizing variables. On the other hand, we have introduced a novel Markov chain Monte Carlo method that uses our recently developed algorithmic techniques (reported by the CO-ALCO group) to significantly improve the reliability of the network discovery results; we have also studied the robustness of such analyses in the presence of onobserved variables. These methodological results were published at the UAI'11, UAI'12, and ECML-PKDD'11 conferences, respectively [2, 3, 4].

Future plans

The Phenomics group was merged to the CO-ALCO group in the beginning of 2013.

Societal, economical and technical impact

The data analysis results have direct impact on the hypothesis formation by the domain experts. We expect this further lead to new studies and knowledge that have impact on practices relevant for public health.

Our contributions to data analysis methods more generally are expected to have impact on data analysis in other domains and on further development on the methods. In addition, the methodological issues raise research problems that motivate and steer the research on algorithm theory, especially in our CO-ALCO group.

Cooperation

Within Algodan, the group works in close collaboration with the CO-ALCO group, partly because of the shared researchers (Koivisto and Parviainen). The group was merged to CO-ALCO in the beginning of 2013.

National and international collaborations:

- Center for Neurobehavioral Genetics at the University of California Los Angeles (UCLA), USA; analysis of genotype and phenotype data; joint publications.
- Institute for Molecular Medicine in Finland (FIMM) and National Institute of Health and Welfare (THL); analysis of genotype and phenotype data; joint publications.
- Departments of Psychiatry and of Public Health and General Practice at University of Oulu, Finland; analysis of the NFBC66 data; joint publications.

Selected publications

1. Eliza Congdon, Susan Service, Jaana Wessman, Jouni K. Seppänen, Stefan Schönauer, Jouko Miettunen, Hannu Turunen, Markku Koiranen, Matti Joukamaa, Marjo-Riitta Järvelin, Leena Peltonen, Juha Veijola, Heikki Mannila, Tiina Paunio, and Nelson B. Freimer. Early environment and neurobehavioral development predict adult temperament clusters. PloS One 7(7), 2012.
2. Teppo Niinimaki and Pekka Parviainen. Local structure discovery in Bayesian networks. 28th Conference on Uncertainty in Artificial Intelligence (UAI 2012), pp. 634-643, AUAI, 2012.
3. Teppo Niinimäki, Pekka Parviainen, and Mikko Koivisto. Partial order MCMC for structure discovery in Bayesian networks. 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011), pp. 557-564, AUAI, 2011.
4. Pekka Parviainen and Mikko Koivisto. Ancestor relations in the presence of unobserved variables. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2011), LNCS 6912, pp. 581-596, Springer, 2011.
5. Jaana Wessman, Stefan Schönauer, Jouko Miettunen, Hannu Turunen, Pekka Parviainen, Jouni K. Seppänen, Eliza Congdon, Susan Service, Markku Koiranen, Jesper Ekelund, Jaana Laitinen, Anja Taanila, Tuija Tammelin, Mirka Hintsanen, Laura Pulkki-Råback, Liisa Keltikangas-Järvinen, Jorma Viikari, Olli T. Raitakari, Matti Joukamaa, Marjo-Riitta Järvelin, Nelson Freimer, Leena Peltonen, Juha Veijola, Heikki Mannila, and Tiina Paunio. Temperament clusters in a normal population: implications for health and disease. PLoS ONE 7(7), 2012.

Team Combinatorial Pattern Matching

Combinatorial pattern matching algorithms and applications

Members

- Esko Ukkonen, Professor, Group leader
- Leena Salmela, Postdoctoral researcher (8/2009-)
- Emanuele Giaquinta, Postdoctoral researcher (2/2012-)
- Simon Puglisi, Postdoctoral researcher (1/2012-)
- Jarkko Toivonen, Doctoral student
- Otto Solin, Doctoral student
- Dominik Kempa, Doctoral student
- Antti Laaksonen, Doctoral student
- Johannes Ylinen, MSc student (2010–2012)

Mission of the group

The group develops theoretical concepts, models, and algorithms for sequence-related problems from biological sequence analysis and other areas. The algorithm-theoretic research is complemented by application-oriented work which is done in close collaboration with many groups of biologists who provide up-to-date problems and new data to be analyzed using the new methods developed in the group.

Research activities

S - Sequence analysis

Prediction of gene regulatory motifs in DNA and de novo sequencing has been the major applied research topics in the group.

The activity of genes is regulated by so-called regulatory modules that are complexes of proteins called transcription factors (TF). Such complexes are formed by TFs that bind to each other and to the DNA on specific TF binding sites. The position weight matrix (PWMs) is the standard probabilistic model for the binding affinity between a transcription factor (TF) and DNA. Accurate PWM models are an essential component of gene regulation models. Our collaborator Jussi Taipale (Univ Helsinki and Karolinska Institute, Sweden) has developed a novel technology, based on the so-called SELEX procedure, for high-throughput sampling of the binding sites of all TFs and of pairs of TFs. SELEX yields very large training data for learning PWMs and more advanced models. We have developed a novel 'multinomial' learning algorithm for PWMs that gives accurate models [2, 14]. We have also developed learning algorithms for finding more accurate models that consist of multiple PWMs for the same TF. The binding affinity of a TF may have several local peaks within the possible binding sites. Each peak has its characteristic sequence that we use as a seed to construct a PWM for that local maximum [22]. Moreover, we have developed new co-operative binding models and learning algorithms for them, to model binding sites of TF complexes that consist of two factors (dimers) [14]. Such dimeric models will be used as building blocks for a general Markov chain model that we are currently developing for chains of binding sites that constitute a putative regulatory module.

In de novo sequencing, the group participated in Professor Ilkka Hanski's (Metapopulation Research Unit, Department of Ecology and Evolutionary Biology) major project of sequencing of the entire genome of the Glanville fritillary butterfly (Melitaea cinxia); this is the first eukaryote sequencing project in Finland. The genome has been sequenced, assembled and annotated [23]. The assembled draft genome consists of

8.262 scaffolds of total length 390 Mbp. The genome has been compared to other published butterfly genomes and an exceptionally high level of synteny is found in the gene order between the species. The Granville fritillary is the first sequenced butterfly with the ancient karyotype of 31 which also allows us to find patterns in the chromosome fusion events that have occurred in other species. We have developed the data analysis pipeline for high-throughput sequencing, with novel algorithms for error correction [8] and scaffolding [5]. For the most time-consuming step of finding overlaps between sequence segments we use the very fast compressed suffix-array implementation developed by Veli Mäkinen's group. A new metric, normalized N50, and a method for evaluating it is proposed for the quality of a set of contigs or scaffolds produced by a genome assembler [11].

In basic research in algorithms for sequences, the group has produced several results. For example, we proved a practical result about the optimality of a class of string searching algorithms in the average case [9]. We also introduced an alphabet sampling technique to speed up string searching algorithms [10]. In [17, 19] we presented a novel method and fast algorithms to find, in a given DNA sequence, the binding sites of a TF motif described using a generalized Position Weight Matrix that also models dependencies between the motif positions.

D - Discovery of hidden structure in high-dimensional data

In a joint project with the Division of Astronomy of the university (Doc Lauri Haikala), we applied Gaussian mixture modeling to locate stellar clusters (potential formation areas of new stars) in the recent data of the United Kingdom Infrared Telescope Infrared Deep Sky Survey. Our search found 137 previously unknown cluster candidates and 30 previously unknown sites of star formation [6]. In the Vía Láctea Survey (VVV) catalogue data the search located 88 previously unknown candidates, most of which are embedded stellar cluster candidates, and 39 previously unknown sites of star formation [7].

Future plans

We are working on a few projects whose current status and future plans are as follows:

- Analysis of SELEX data to synthesize models for DNA binding sites of transcription factor complexes (such as dimers); joint work with Jussi Taipale's group. This work will be utilized in a joint EU-funded project with Jussi Taipale and Lauri Aaltonen on systems biology of colorectal cancer. Our role is to develop computational tools for modeling gene regulatory relations and disorders in them. Several yet unpublished results have been obtained.

- The genome project of the Glanville fritillary butterfly with Ilkka Hanski has almost been finished. The final publications are under writing or review.

- We have started with Acad. Prof. Jukka Jernvall a joint work on the evolution of the gene regulatory structures. We utilize a multiple genome variant of our earlier analysis method EEL (Enhances Element Locator) to find conserved regulatory patterns and their evolutionary relations.

- The research on basic algorithmics on strings will be continued very actively. The Algodan CoE has raised a flourishing community of string algorithmics researchers working at the Department of Computer Science. They are expected to have an internationally leading role in the development of this field.

Societal, economical and technical impact

Some of our new algorithms (for the genome assembly [5, 8] or for learning PWMs and their generalizations [14, 22], for example) have potential to become into wide use as they offer improved performance. On application side, for example the new PWM models that are more accurate than the earlier ones [14, 22] have potential for a significant impact in their fields. The new butterfly genome [23] is a major achievement of basic research in its field and may have a strong impact on future research. Some our papers [1, 2, 14] are attracting a rapidly growing number of citations.

Cooperation

Cooperation within Algodan: Collaboration and joint publications with Petteri Kaski, Heikki Mannila, and Veli Mäkinen.

Cooperation within University of Helsinki: Joint ongoing or planned projects with professors Lauri Aaltonen (Biomedicum), Ilkka Hanski (Department of Ecology and Evolutionary Biology), Jussi Taipale (Biomedicum & Karolinska Institutet, Sweden), Jukka Jernvall (Institute of Biotechnology).

International cooperation:

- Professor Alberto Apostolico & Dr. Cinzia Pizzi, University of Padua, Italy and Georgia Tech, USA; analysis of motifs in strings; joint publications.

- Collaborations within EU project SYSCOL with several European and US partners; joint publications.

Selected publications

1. A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G.Wei, M. Enge, M. Taipale, J.M. Vaquerizas, J. Yan, M.J. Sillanpää, M. Bonke, K. Palin, S. Talukder, T.R. Hughes, N.M. Luscombe, E. Ukkonen & J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* 20, 6 (June 2010), 861–873.
2. Gong-Hong Wei, Gwenael Badis, Michael F Berger, Teemu Kivioja, Kimmo Palin, Martin Enge, Martin Bonke, Arttu Jolma, Markku Varjosalo, Andrew R Gehrke, Jian Yan, Shaheynoor Talukder, Mikko Turunen, Mikko Taipale, Hendrik G Stunnenberg, Esko Ukkonen, Timothy R Hughes, Martha L Bulyk and Jussi Taipale. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO Journal* 29 (2010), 2147–2160.
3. A. Pizzi, P. Rastas & E. Ukkonen. Finding significant matches of position weight matrices in linear time. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8, 1 (2011), 69–79.
4. Apostolico, C. Pizzi & E. Ukkonen. Efficient Algorithms for the Discovery of Gapped Factors. Algorithms for Molecular Biology 6:5 (2011), 10 pages.
5. L. Salmela, V. Mäkinen, N. Välimäki, J. Ylinen & E. Ukkonen. Fast scaffolding with small independent mixed integer programs. Bioinformatics 27, 23 (2011), 3259–3265.
6. O. Solin, E. Ukkonen & L. Haikala. Mining the UKIDSS GPS: star formation and embedded clusters. Astronomy & Astrophysics 542, A3 (2012).
7. O. Solin, E. Ukkonen & L. Haikala. Mining the VVV: star formation and embedded clusters. Astronomy & Astrophysics 562, A115 (2014).
8. L. Salmela & J. Schröder. Correcting errors in short reads by multiple alignments. Bioinformatics 27(11):1455-1461 (2011).
9. L. Salmela. Average complexity of backward q-gram string matching algorithms. Information Processing Letters, Volume 112(11):433-437 (2012).

10. F. Claude, G. Navarro, H. Peltola, L. Salmela and J. Tarhio: String matching with alphabet sampling. Journal of Discrete Algorithms, Volume 11 (2012), 37–50.
11. V. Mäkinen, L. Salmela, and J. Ylinen. Normalized N50 assembly metric using gap-restricted co-linear chaining. BMC Bioinformatics, Volume 13 (2012), 255.
12. F. Claude, G. Navarro, H. Peltola, L. Salmela & J. Tarhio. String matching with alphabet sampling. Journal of Discrete Algorithms 11:37-50 (2012).
13. J. Fischer, T. Gagie, T. Kopelowitz, M. Lewenstein, V. Mäkinen, L. Salmela, and N. Välimäki: Forbidden patterns. In Proc. LATIN 2012, Lecture Notes in Computer Science 7256, Springer 2012, 327–337.
14. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G.H., Palin, K., Vaquerizas, J.M., Vincentelli, R., Luscombe, N.M., Hughes, T.R., Lemaire, P., Ukkonen, E., Kivioja, T., Taipale, J.: Dna-binding specifities of human transcription factors. *Cell* 152, 1-2: 327-339 (2013).
15. Emanuele Giaquinta and Szymon Grabowski. New algorithms for binary jumbled pattern matching. *Information Processing Letters*, 113(14-16): 538–542, 2013.
16. Emanuele Giaquinta, Szymon Grabowski, and Kimmo Fredriksson. Approximate pattern matching with k-mismatches in packed text. *Information Processing Letters*, 113(19-21):693–697, 2013.
17. E. Giaquinta, S. Grabowski, and E. Ukkonen. Fast matching of transcription factor motifs using generalized position weight matrix models. *Journal of Computational Biology*, 20(9):621–630, 2013.
18. Ferdinando Cicalese, Travis Gagie, Emanuele Giaquinta, Eduardo Sany Laber, Zsuzsanna Lipt´ak, Romeo Rizzi, and Alexandru I. Tomescu. Indexes for jumbled pattern matching in strings, trees and graphs. In Proc. SPIRE, Lecture Notes in Computer Science 8214, Springer 2013, 56–63.
19. E. Giaquinta, K. Fredriksson, S. Grabowski, and E. Ukkonen. Motif matching using gapped patterns. In Proc. IWOCA 2013, Lecture Notes in Computer Science 8288, Springer 2013, 448–452.
20. Domenico Cantone, Simone Faro, and Emanuele Giaquinta. Text searching allowing for inversions and translocations of factors. *Discrete Applied Mathematics*, 163:247–257, 2014.
21. Kimmo Fredriksson and Emanuele Giaquinta. On a compact encoding of the swap automaton. *Information Processing Letters*, 2014. To appear.
22. Kazuhiro R. Nitta, Arttu Jolma, Teemu Kivioja, Junaid Akhtar, Korneel Hens, Bart Deplancke, Eileen Furlong, and Jussi Taipale: Conservation and divergence of transcription factor binding specificity. Submitted manuscript.
23. V. Ahola, R. Lehtonen, P. Somervuo, L. Salmela, P. Koskinen, P.Rastas, N. Välimäki, J. Kvist, L. Paulin, N. Wahlberg, M. Taipale, S. Luo, L. C. Ferguson, J. Tanskanen, R. Waterhouse, M. A. de Jong, A. Duplouy, O.-P. Smolander, H. Vogel, Z. Cao, R. C. McCoy, K. Qian, E. A. Hornett, W. S. Chong, Q. Zhang, F. Ahmad, J. K. Haukka, A. Joshi, J. Salojärvi, C. W. Wheat, E. Grosse-Wilde, M. Turunen, A. Vähärautio, E. Ukkonen, D. Huges, D. Lawson, V. Mäkinen, M. Goldsmith, L. Holm, M. Frilander, P. Auvinen, I. Hanski. Glanville fritillary genome retains ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat. Gen.* (submitted).

Succinct Data Structures (SuDS)

Members

- Veli Mäkinen, Professor, Group leader (-2011)
- Niko Välimäki, Doctoral student (-2012)
- Jouni Sirén, Doctoral student (-2012)
- Santeri Pietilä, Research Assistant (3/2011-8/2011)

Mission of the group

The study of succinct data structures extends traditional data compression with the functionality preserving property: data structure functions need to be efficiently computable directly from the compressed representation. In addition to providing and analyzing new succinct data structures, the group contributes by engineering open source implementations targeted to applications especially in biological sequence analysis and information retrieval.

Research activities

S - Sequence analysis, S.1 String algorithms /
F - Foundations of algorithmic data analysis, F.1 Theory of string matching

Compressed representations for highly repetitive sequence collections, such as version histories and collections of genomes of individuals within same species, are developed in [1]. This extensive study includes combinations of static cases, dynamic cases, different models to measure high repetitiveness, tradeoffs, and extensions to suffix tree representation. This is the first study beyond the familiar k-th order model in compressed text indexes. In [2] we propose, implement, and experiment a compressed solution for XML indexing. The solution supports XPath query language together with full-text predicates such as prefix, suffix, contains, less-than, etc. In principle, the solution is a carefully designed merge of existing solutions from compressed tree representations and compressed text representations, but it also contains new insights into XPath query evaluation. On an existing benchmark, the new index is faster on all queries than its competitors. Space requirement is better or similar to its competitors. We have recently added the support for XML documents representing a genome annotation database, enabling queries by annotation restrictions (e.g. organism type, gene function, promoter, etc.) and sequence content (PWM matrix and approximate search support). This work with some other improvements by co-authors is now submitted to a journal.

In [3] we extended our previous results on exact substring searches to approximate search, giving some new insights into the timely DNA sequencing read alignment problem. Then we worked on new solutions to the classical de novo fragment assembly problem using our new scalable approach to approximate overlap alignment [4]. Jouni Sirén continued his work on compressed suffix arrays, extending the ideas to longest common prefix arrays [3]. Then later he got in contact with Paolo Ferragina and Rossano Venturini to develop improved suffix array sampling scenarios [7]. Our most recent developments include an extension of Burrows-Wheeler transform to finite automaton representing reference genome together with its common variations among the population [6]. This enables a space-efficient index structure to be constructed to support efficient DNA sequencing read alignment to a rich model of the population.

Future plans

Starting from 2012, the group is no longer part of ALGODAN. See below cooperation section for the reason.

Societal, economical and technical impact

Our new developments in [6] are very useful in the variation calling application and led to the collaboration on cancer genetics research (see below).

Cooperation

Cooperation within Algodan: Collaboration with the group of Esko Ukkonen on and de novo fragment assembly, with the group of Juha Kärkkäinen on text index construction algorithms, and with Petteri Kaski on space-efficient traversals on huge implicit graphs.

Cooperation within University of Helsinki: Starting from 2012 the group moved to the new Center of Excellence in Cancer Genetics Research, changing the name to genome-scale algorithmics. The new center is led by Professor Lauri Aaltonen (Biomedicum), with whom we had earlier joint work in the analysis of next-generation sequencing data (Riku Katainen from our group moved to Aaltonen's group in 2010). Then we work also with Professor Ilkka Hanski (Department of Ecology and Evolutionary Biology) on the assembly of the genome of Glanville fritillary butterfly (Melitaea cinxia).

International cooperation (during 2010-2012):

- Professor Gonzalo Navarro, University of Chile, Theory of string matching, joint publications, software development, exchange of researchers
- Dr. Johannes Fischer, Karlsruhe Institute of Technology, String mining & compressed suffix trees, joint publications, software development, exchange of visits
- Senior Researcher Sebastian Maneth, NICTA Kensington Research Lab, Sydney, Australia, XML indexing, joint publications, software development, exchange of visits
- Professor Paolo Ferragina, University of Pisa, Theory of string matching, joint publications, exchange of researchers

Selected publications

1. Veli Mäkinen, Gonzalo Navarro, Jouni Sirén, and Niko Välimäki. Storage and Retrieval of Highly Repetitive Sequence Collections. Journal of Computational Biology. March 2010, 17(3): 281-308. Earlier in RECOMB 2009 & SPIRE 2008.
2. Diego Arroyuelo, Francisco Claude, Sebastian Maneth, Veli Mäkinen, Gonzalo Navarro, Kim Nguyen, Jouni Sirén, and Niko Välimäki. Fast In-Memory XPath Search using Compressed Indexes. In Proc. 26th IEEE International Conference on Data Engineering (ICDE 2010), March 1-6, 2010, Long Beach, California, USA.
3. Veli Mäkinen, Niko Välimäki, Antti Laaksonen, and Riku Katainen. Unified View of Backward Backtracking in Short Read Mapping. In Ukkonen Festschrift 2010 (Eds. Tapio Elomaa, Pekka Orponen, Heikki Mannila), Springer-Verlag, LNCS 6060, pp. 182-195, 2010.
4. Jouni Sirén. Sampled Longest Common Prefix Array. In Proc. CPM 2010, Springer LNCS 6129, pp. 227-237, New York, USA, June 21-23, 2010.
5. Niko Välimäki, Susana Ladra, and Veli Mäkinen. Approximate All-Pairs Suffix/Prefix Overlaps. In Proc. 21st Annual Symposium on Combinatorial Pattern Matching (CPM 2010), Springer-Verlag, LNCS 6129, pp. 76-87, New York, USA, 21-23 June 2010.
6. Jouni Sirén, Niko Välimäki, and Veli Mäkinen. Indexing Finite Language Representation of Population Genotypes. In Proc. Algorithms in Bioinformatics (WABI 2011), Springer-Verlag, LNCS 6833, pp. 270-281, Saarbrücken, Germany, September 5-7, 2011.
7. Paolo Ferragina, Jouni Sirén, and Rossano Venturini. Distribution-Aware Compressed Full-Text Indexes. In Proc. 19th Annual European Symposium (ESA 2011), Springer-Verlag, LNCS 6942, pp. 760-771, Saarbrücken, Germany, September 5-7, 2011.

Practical Algorithms and Data Structures on Strings (PADS)

Members

- Juha Kärkkäinen, University Researcher, Group leader
- Simon Puglisi, Postdoctoral researcher (4/2012-)
- Dominik Kempa, Doctoral student (7/2011-)
- Pekka Mikkola, MSc student (5/2010 – 8/2012)

Mission of the group

The group develops fast and practical algorithms and data structures for fundamental problems arising in sequence analysis. The research is based on thorough understanding of both the combinatorial properties of the problems and the properties of modern computers. The goal is not only to obtain better algorithms but to understand why they are better.

Research activities

S – Sequence analysis
F – Foundations of algorithmic data analysis

When indexing and analysing massive amounts of sequential data, conventional data structures such as the suffix tree are increasingly being replaced by text indexes based on compressible representations of sequences, particularly the Burrows-Wheeler transform (BWT) and the Lempel-Ziv factorisation (LZF), and context free grammars (CFG). We have studied several aspects related to these new indexes.

1. Index construction. We have developed several new algorithms for computing the LZF that are simultaneously faster and more space efficient than previous algorithms [4,8,9]. To deal with ever larger amounts of data, we have recently designed and implemented practical external memory algorithms for computing the LZF [16] as well as the suffix array [17] and the LCP array [19], which are key components in constructing BWT-based text indexes. Of more theoretical interest are in-place algorithms for computing the BWT and its inverse [6].

Some of the subproblems arising in text index construction are interesting in their own right. We have indentified and studied two new string matching variants, longest prefix matching [10] and string range matching [18]. Algorithms developed in these studies are used as a part of the index construction algorithms mentioned above. We have also achieved a big theoretical improvement to the problem of sparse suffix sorting [14] that arises in some suffix array construction algorithms.

2. Index components. The FM-index is the best known BWT-based text index and we have developed new techniques for implementing the basic components of the FM-index reducing both the space requirement and the query time [1, 15].

3. Bidirectional indexes. We have developed compressed text indexes that support extending and contracting the pattern from both ends during a search, which allows a much richer set of pattern matching and pattern discovery operations on the index [11, 13].

4. Document retrieval. We have developed techniques for using compressed text indexes for document retrieval, which is one of the most important, complex and demanding application of text indexes [7,12].

5. Data compression. We have also developed practical techniques for pure (non-indexing) data compression [2, 3, 5], some of which are included in an experimental, open source compressor available at https://github.com/pjmikkol/bwtc.

Future plans

The group continues to study all aspects of compressed text indexes but particular focus areas are practical techniques for index construction and the basic components of the indexes. For index construction the next direction is handling ever larger amounts of data using external memory, parallel and distributed computing. For basic components, the focus will be on highly repetitive data, where the difficulty is achieving good compression and fast queries simultaneously.

Societal, economical and technical impact

Many of the algorithms and techniques developed by the group are simple and practical and have the potential for being included in many applications. When dealing with massive data, faster and more space efficient techniques can provide substantial economic benefits too.

Cooperation

Cooperation within Algodan and the University of Helsinki:

We have worked on bidirectional indexes and information retrieval with the group of Veli Mäkinen, which left Algodan in 2012 to join the Center of Excellence in Cancer Genetics Research.

National and international cooperation:

- Kalle Karhu, Aalto University, Finland; bidirectional indexes and document retrieval, joint publications.
- Maxime Crochemore and Costas Iliopoulos, King's Collage London, UK; BWT computation; joint publication, research visits.
- Hideo Bannai and Tomohiro I, Kyushu University, Japan; sparse suffix sorting; joint publication, research visits.
- Simon Gog, University of Melbourne, Australia; bidirectional indexes; joint publication, research visit.
- Gonzalo Navarro and Jouni Siren, University of Chile; document retrieval; joint publications.
- Roberto Grossi, University of Pisa, Italy; BWT computation; joint publication.
- Gad M. Landau, University of Haifa, Israel: BWT computation; joint publication.
- German Tischler, The Wellcome Trust Sanger Institute, UK; data compression; joint publication.
- Peter Sanders, Karlsruhe Institute of Technology, Germany; research visit.

Selected publications

1. Juha Kärkkäinen, Simon J. Puglisi. Fixed Block Compression Boosting in FM Indexes. In Proc. 18th Symposium on String Processing and Information Retrieval (SPIRE 2011), Springer, 2011, pp. 174-184.
2. Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. Slashing the Time for BWT Inversion. In Proc. 2012 Data Compression Conference (DCC 2012), IEEE Computer Society, 2012, pp. 99-108.
3. Juha Kärkkäinen, Pekka Mikkola, Dominik Kempa. Grammar Precompression Speeds Up Burrows-Wheeler Compression. In Proc. 19th Symposium on String Processing and Information Retrieval (SPIRE 2012), Springer, 2012, pp. 330-335.

4. Dominik Kempa, Simon J. Puglisi. Lempel-Ziv factorization: Simple, fast, practical. In Proc. Meeting on Algorithm Engineering & Experiments (ALENEX 2013), SIAM, 2013, pp. 103-112.
5. Juha Kärkkäinen, German Tischler. Near in Place Linear Time Minimum Redundancy Coding. In Proc. 2013 Data Compression Conference (DCC 2013), IEEE Computer Society 2013, pp. 411-420.
6. Maxime Crochemore, Roberto Grossi, Juha Kärkkäinen, Gad M. Landau. A Constant-Space Comparison-Based Algorithm for Computing the Burrows-Wheeler Transform. In Proc. 24th Symposium on Combinatorial Pattern Matching (CPM 2013), Springer, 2013, pp. 74-82.
7. Travis Gagie, Kalle Karhu, Gonzalo Navarro, Simon J. Puglisi, Jouni Sirén. Document listing on repetitive collections. In Proc. 24th Symposium on Combinatorial Pattern Matching (CPM 2013), pp. 107-119, 2013
8. Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. Linear Time Lempel-Ziv Factorization: Simple, Fast, Small. In Proc. 24th Symposium on Combinatorial Pattern Matching (CPM 2013), Springer, 2013, pp. 189-200.
9. Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. Lightweight Lempel-Ziv Parsing. In Proc. 12th Symposium on Experimental Algorithms (SEA 2013), Springer, 2013, pp. 139-150.
10. Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. Crochemore's String Matching Algorithm: Simplification, Extensions, Applications. In Proc. Prague Stringology Conference (PSC 2013), Czech Technical University in Prague, Czech Republic, 2013, pp. 168-175.
11. Djamal Belazzougui, Fabio Cunial, Juha Kärkkäinen, Veli Mäkinen. Versatile Succinct Representations of the Bidirectional Burrows-Wheeler Transform. In Proc. 21st European Symposium on Algorithm (ESA 2013), Springer, 2013, pp. 133-144.
12. Travis Gagie, Juha Kärkkäinen, Gonzalo Navarro, Simon J. Puglisi. Colored range queries and document retrieval. Theoretical Computer Science 483, pp. 36-50, 2013.
13. Simon Gog, Kalle Karhu, Juha Kärkkäinen, Veli Mäkinen, Niko Välimäki. Multi-pattern matching with bidirectional indexes. Journal of Discrete Algorithms 24, pp. 26-39, 2014.
14. Tomohiro I, Juha Kärkkäinen, Dominik Kempa. Faster Sparse Suffix Sorting. In Proc. 31st Symposium on Theoretical Aspects of Computer Science (STACS 2014), Springer, 2014, to appear.
15. Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. Hybrid Compression of Bitvectors for the FM-Index. In Proc. 2014 Data Compression Conference (DCC 2014), IEEE Computer Society 2014, to appear.
16. Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. Lempel-Ziv Parsing in External Memory. In Proc. 2014 Data Compression Conference (DCC 2014), IEEE Computer Society 2014, to appear.
17. Juha Kärkkäinen, Dominik Kempa. Engineering a Lightweight External Memory Suffix Array Construction Algorithm. In Proc. 2nd Conference on Algorithms for Big Data (ICABD 2014), to appear.
18. Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. String range matching. In Proc. 25th Symposium on Combinatorial Pattern Matching (CPM 2014), Springer, 2014, to appear.
19. Juha Kärkkäinen, Dominik Kempa. LCP Array Construction in External Memory. Submitted to a conference, 2014.

Computational Geometry

Members

- Valentin Polishchuk, Academy Postdoctoral Researcher (1/2011-12/2013), Docent (4/2012-), Group leader
- Mikko Nikkilä, MSc student (3/2012-)
- Mikko Sysikaski, MSc student (6/2010-9/2013)
- Topi Talvitie, MSc student (6/2013-)
- Juha-Antti Isojärvi, MSc student (-2012)
- Sylvester David Eriksson-Bique, Doctoral student (-2012)

Mission of the group

Geometric data analysis, visualization and processing are inherent to numerous domains ranging from motion planning to VLSI to geographic information systems to robotics. We design, analyze, and implement computational-geometry algorithms applicable to current and future tasks in intelligent path design, cartography, shape reconstruction and sensor networks.

Research activities

F - Foundations of algorithmic data analysis

We extended classical algorithmic and combinatorial results from discrete network flows to continuous geometric domains; we will continue investigations into the mapping between discrete graph notions and their continuous analogues in geometry, motivated by motion planning and coordination challenges arising in air traffic industry. Our most recent results were presentred at ACM SIGSPATIAL GIS'13 [1] and are to appear in ACM/SIGGRAPH Symposium on Copmutational Geometry 2014 [2].

In our cross-disciplinary work bridging shape modeling and geophysics, we developed novel shape reconstruction methods robust to noise and outliers; our first paper on application of the tools to seismic data analysis has just been accepted to Geophysical Journal International [3]. Our other recent publications on the theme include a paper [4] in IEEE Transactions on Visualization and Computer Graphics -- a top journal in the field.

Sensor network is a source of a bulk of measurements taken by the sensors over time; we look at a variety of computational-geometry questions that arise from processing sensor-network data. Our publications in the area include a paper [5] in MobiHoc -- a top conference in the field.

Future plans

We will continue research in geometric methods for a variety of applications. On the motion planning frontier, we will investigate further fundamental geometric problems of planning paths under a multitude of constraints and requirements. Research on shape approximation will lead to automatic shape reconstruction approaches for seismic data analysis. For sensor networks we will develop efficient data gathering methods.

Societal, economical and technical impact

The research on air traffic motion coordination provides decision-support tools for air traffic industry's humans-in-the-loop – traffic controllers, dispatchers, managers; given the amount of the world air traffic, even small improvements to the current procedures, even implemented on a local scale, lead to huge savings in operating costs, to decrease in the environmental impact of air traffic, and to increased safety

and efficiency of flight management. The high level of theoretical abstraction pertinent to our algorithmic work allows one to use our results also in other domains – nanostrucutre design, crowd evacuation, robotics, computer games. Efficient processing of geospatial data representing real-world terrains may enable faster and less costly search-and-rescue operations.

Analyzing geophysical data ultimately leads to better prediction of seismic events. In general, shape approximation and simplification tools are applicable in motion planning, object recognition and data analysis.

Sensor networks are in use for surveillance, monitoring and tracking of objects of very different types – from wildlife to goods in a warehouse; improved algorithms for the networks imply savings in the network management and data handling.

Cooperation

Cooperation within the universities

- Our group is part of the New Paradigms in Computing group at HIIT

International cooperation:

Our group collaborates with researchers all around the world, coauthoring both with academia (Carnegie Mellon, Stony Brook, University of Arizona, Institute of Dynamics of Geospheres, many European colleagues) and industry (IBM, Google, Mathworks, Metron Aviation). We regularly visit our colleagues and host visitors in Helsinki.

Participation in European research networks

- ComplexWorld.eu. Mastering system complexity.
- Toward Higher Levels of Automation in ATM.

Selected publications:

1. A. Efrat, M. Nikkilä, V. Polishchuk. Sweeping a Terrain by Collaborative Aerial Vehicles. ACM SIGSPATIAL GIS'13.
2. S. Eriksson-Bique, V. Polishchuk, M. Sysikaski. Optimal geometric flows via dual programs. To appear in ACM/SIGGRAPH SoCG'14.
3. M. Nikkilä, V. Polishchuk, D. Krasnoshchekov. Robust estimation of seismic coda shape. To appear in Geophysical Journal International.
4. P. Bak, E. Packer, H. Ship, M. Nikkilä, V. Polishchuk. Visual Analytics for Spatial Clustering: Using a Heuristic Approach for Guided Exploration. Special issue of IEEE Transactions on Visualization and Computer Graphics on IEEE VIS (VAST)'13.
5. S. Sankararaman, K. Abu-Affash, A. Efrat, S. Eriksson-Bique, V. Polishchuk, S. Ramasubramanian, M. Segal. Optimization Schemes for Protective Jamming. ACM MobiHoc'12.

C-BRAHMS Group

Members

- Kjell Lemström, PhD, Docent, Group leader
- Teppo Ahonen, Doctoral student
- Antti Laaksonen, Doctoral student
- Mika Laitinen, MSc student (2010–2011)
- Simo Linkola, MSc student (2011)
- Lari Rasku, MSc student (2012–2013)

Mission of the group

The C-BRAHMS project aims at designing and developing efficient methods for computational problems arising from music comparison, retrieval, and analysis. The main concentration is on retrieving polyphonic music in large-scale music databases containing symbolically encoded music. The project utilizes various algorithmic techniques together with findings in musicology and music psychology to achieve efficient, musically meaningful results.

Research activities

S: Sequence analysis

We have studied and developed new algorithms for analyzing, classifying and retrieving musical sequences. In order to find efficient and effective tools for various tasks, we have used a variety of different modelings of music, similarity measures and methods, with the aim of finding the optimal combination for the given task. For finding occurrences or recurrences in symbolically-encoded polyphonic music, computational-geometry-based techniques seem very promising: we have extended the setting behind our original sweep-line music-retrieval algorithm [1] to new musical point-pattern matching problems [2, 3, 10, 11] and have also adapted the framework of mathematical morphology to this problem area [7]. Most recently, we have adapted the method to work directly with audio music [16].

Another research branch of the group is in applying normalized compression distance (NCD) in content-based music retrieval (CBMR) tasks. In his PhD project, MSc Ahonen have focused on CBMR using NCD both in audio [4, 5, 11] and in symbolic [6] domains. In addition to studying how the musical information should be represented for compression-based similarity measuring, the work has presented several novel ideas for extending the pairwise NCD similarity measuring to sets of objects [5] and explored how NCD can be used in classification and clustering instead of more commonly used distance metrics [6,11]. Recently, we have also suggested an adaptive representation for music features that both allows efficient similarity measuring with any common similarity metric while preserving essential distinguishing power [12].

A new research branch of the group is automatic music transcription from a musician's viewpoint. In his PhD project, MSc Laaksonen, currently concentrates on chord transcription with the aim of finding new ways to utilize the musical context in the transcription. His latest findings suggest that the melody context, which has not been used in chord transcription so far, plays an important role in cases where the pure chord content is ambiguous. Laaksonen aims at creating transcriptions which sound good as a whole in a real musical performance rather than creating exact transcriptions from individual chords. Results of this project have been reported in publications [13-15].

Future plans

Dr. Lemström is co-editing a text-book (working title "The Oxford Handbook of Automated Knowledge Discovery in Music") with two internationally leading scientists in the field, Professor Geraint Wiggins and Professor Roger Dannenberg.

MSc Teppo Ahonen's PhD thesis, *Compression-based Similarity Measuring for Practical Applications in Content-based Music Information Retrieval*, is set to be defended in fall 2012. The thesis focuses on using NCD and other compression-based similarity measures to measure similarity between tonal features extracted from music data. The thesis provides insight into (1) what features are essential when measuring tonal similarity between pieces of music, (2) how the features should be represented for a compression-based similarity metric, (3) what are the advantages and disadvantages of using NCD for measuring tonal similarity, and (4) how the methodology can be applied for retrieval and machine learning tasks.

Cooperation

- Gerant Wiggins, Goldsmiths College, University of London, UK, visits, joint book project
- Roger Dannenberg, Carnegie Mellon University, USA, joint book project
- David Rizo, Jose Manuel Iñesta, University of Alicante, joint publications
- David Meredith, Aalborg University, visits, joint publications

Selected publications

1. Esko Ukkonen, Kjell Lemström and Veli Mäkinen: Geometric Algorithms for Transposition Invariant Content-Based Music Retrieval. In Proc. ISMIR'03 4th International Conference on Music Information Retrieval, Baltimore, October 26-30, 2003, pp. 193-199.
2. Kjell Lemström, Niko Mikkilä and Veli Mäkinen: Filtering Methods for Content-Based Retrieval on Indexed Symbolic Music Databases. In Journal of Information Retrieval, 13 (1), pp. 1-21, 2010.
3. Kjell Lemström: Transposition and Time-Scale Invariant Geometric Music Retrieval. In Essays Dedicated to Esko Ukkonen on the Occasion of His 60th Birthday (Eds. Tapio Elomaa, Pekka Orponen, and Heikki Mannila), Springer- Verlag, LNCS 6060, 2010.
4. Teppo E. Ahonen: Combining Chroma Features for Cover Version Identification. Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010) Utrecht, The Netherlands, August 2010, pp. 165-170.
5. Teppo E. Ahonen: Compressing Lists for Audio Classification. Proceedings of the 3rd International Workshop on Machine Learning and Music (MML 2010), Florence, Italy, October 2010.
6. Teppo E. Ahonen, Kjell Lemström, Simo Linkola: Compression-based Similarity Measures in Symbolic, Polyphonic Music. Proceedings of the 12th International Society for Music Information. Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR, 11, Miami, USA, October 2011, pp. 91-96.
7. Mikko Karvonen, Mika Laitinen and Kjell Lemström: Error-tolerant Content-based Music-retrieval with Mathematical Morphology. LNCS 6684 Springer 2011, ISBN 978-3-642-23125-4. pp. 321-337.
8. David Rizo, Kjell Lemström and Jose-Manuel Iñesta: Polyphonic Music Retrieval with Classifier Ensembles. Journal of New Music Research, 40, 4, 2011, pp. 313-325.
9. Kjell Lemström and Mika Laitinen: Transposition and Time-warp Invariant Geometric Music Retrieval Algorithms. Proceedings of the 2011 IEEE International Conference on Multimedia and Expo, Barcelona, Spain, July 2011, pp. 1-6.

10. Mika Laitinen and Kjell Lemström: Dynamic Programming in Transposition and Time-Warp Invariant Polyphonic Content-Based Music Retrieval. Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11), Miami, Florida, USA, October 2011, pp. 369-374.
11. Teppo E. Ahonen: Compression-based Clustering of Chromagram Data: New Method and Representations. To appear in Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR 2012). London, UK, June 2012.
12. Simo Linkola, Lari Rasku and Teppo E. Ahonen: Expandable String Representation for Music Features Using Adaptive-Size Alphabets. Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR 2013), Marseille, France, October 15-18, 2013, pp. 920-927.
13. Antti Laaksonen: Ambiguity in automatic chord transcription: recognizing major and minor chords. Proceedings of the 10th international workshop on Adaptive Multimedia Retrieval, Copenhagen, Denmark, 2012.
14. Antti Laaksonen: Semi-automatic melody extraction using note onset time and pitch information from users. Proceedings of the Sound and Music Computing Conference, Stockholm, Sweden, 2013.
15. Antti Laaksonen: Efficient and simple algorithms for time-scaled and time-warped music search. Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research, Marseilles, France, 2013.
16. Antti Laaksonen and Kjell Lemström: On finding symbolic themes directly from audio using dynamic programming. Proceedings of the 14th International Society for Music Information Retrieval Conference, Curitiba, Brazil, 2013.

Computational Linguistics Group

Members

- Roman Yangarber, University Researcher, Group leader
- Mian Du, Doctoral student
- Lidia Pivovarova, Doctoral student (2011-2012 visiting from St Petersburg State University; at UH 11/2012-)
- Javad Nouri, MSc student
- Guowei Lv, MSc student (MSc 2014)
- Matthew Pierce, MSc student
- Natalia Ostapuk, Doctoral student (2013–2014 visiting from St Petersburg State University)
- Silja Huttunen, Doctoral student (Linguistics) (-2011)
- Hannes Wettig, Doctoral student (- 2012)
- Peter von Etter, MSc, (-2013)
- Arto Vihavainen, MSc student (-2011)
- Suvi Hiltunen, MSc student (-2011)
- Mikhail Novikov, MSc student (-2011)

Mission of the group

The group works on various problems in analysis of linguistic data. We investigate how language conveys information, how information can be extracted from linguistic data, and how hidden, underlying structure can be learned from observed linguistic data. The research programme combines empirical, applied and theoretical approaches to these problems.

Research activities

PULS

The PULS Project builds tools for analysis of plain text, and specifically for surveillance of on-line news media. The group conducts research in the field of information extraction: i.e., identifying pre-defined types of events in text. PULS participates in projects on several different knowledge domains: business intelligence, epidemiological surveillance, and cross-border security.

For example, in the domain of epidemic surveillance the system identifies, in each news article, how many people have been affected with what condition, where, when, etc. The system is operational (at puls.cs.helsinki.fi with access to detailed analysis obtainable from puls@cs.helsinki.fi). The project is developed in collaboration with international partner organisations, who act as research partners and users. A central research theme in the project is automating the acquisition of domain knowledge from plain text. Machine learning methods, especially weakly-supervised learning, are at the core of the methods to bootstrap new systems for analysing documents in new domains quickly and accurately. An important objective of PULS is to investigate and model linguistic phenomena in the context of real-world applications. The key benefit of engaging end-users is that they provide high-quality feedback, as well as annotated data for training and testing, which allows us to experiment on a large scale with supervised and weakly-supervised methods. Prior work deals with text analysis in toy-like laboratory settings, where the amount of data is limited; shortage of data is a bottleneck in NLP research in general, and in news surveillance in particular. PULS has been carefully designed and coordinated over the last 4 years with a view toward obtaining good data in collaboration with real-world end users.

Etymon

The Etymon Project aims to develop computational methods for studying the origin and development of languages and language families, i.e., genetically related groups of languages. We use databases of lexical material from these languages and developing computational methods for studying language relationships. Currently these methods are applicable generally in principle, and are currently being tested on the Uralic, Turkic, Indo-European and Xoisan families. Etymon brings together an international, inter-disciplinary team of researchers, with complementary expertise in the areas of computational, comparative and historical linguistics.

Research on language evolution and inter-linguistic relationships has gone on for over two centuries, during which linguists devised methods for discovering patterns of correspondence, and for testing hypotheses. Because linguistic data is often scarce and incomplete, modern non-computational methods for investigating linguistic relationships leave a large number of "grey areas" – questions that current theory is unable to answer with certainty. In particular, the Uralic data is strikingly uncertain, with conflicting theories in current scholarship. Earlier research on computational etymology over the last decade has focused largely on the Indo-European language family (traditionally at the center of historical-linguistic research) with less emphasis on other families. Etymon has developed novel computational approaches based on the information-theoretic MDL (minimum description length) Principle, to model the etymological correspondences and evolution within the family. One of our aims is to illuminate some of the uncertainties in existing data sets. Another is to compare results from the computational methods with those obtained by traditional, manual methods, directly from the data -- and using all available data in an objective, unbiased fashion. Relationships among words from a group of related languages are captured by discovering regular rules of derivation, from parent to child language, or among sibling languages. The main idea is that the "correct" set of relationships (or a set of rules that efficiently describe the derivation) among a group of related languages will yield a compact encoding for the totality of observable data in these languages.

The key research objectives are:

1. develop novel computational methods for discovering and investigating the relationships, using rich sets of relevant data;
2. apply and test the methods in a wider setting to investigate specific language families, and to the study of relations beyond established language families.

We do not aim to replace the human computational linguist by a model, but rather to aid the linguist by providing additional objective sources of evidence. This will enable one to formulate and quantify evidence that supports the hypotheses and conclusions in ways that were not possible before.

We use data collections provided by partner organisations: – KOTUS (Institute for research on languages of Finland) and the Russian Academy of Sciences.

FinUgRevita

The Computational Linguistics Group has begun a new project (jointly funded by Finland and Hungary) to build novel tutoring and authoring tools for supporting revitalisation of severely endangered languages. The initial focus is on the Uralic languages. There are several difficult research questions. We aim to

a. utilise existing low-level pre-processing resources to provide fundamental automated analysis of arbitrary text in a target language;

b. define a set of states of "competence" -- or "linguistic items" which an expert teacher would consider as helpful for assessing user proficiency,
c. automatically infer a partial order of dependencies among the postulated states -- a dependency network.

The goal is then to create a teaching/tutoring system that is able to use dynamic content, and use the learned dependency network of competence states to create a reliable model of competence of a given user. This is to assure that the user will not be bored with test examples that are too easy, nor discouraged by test examples that are too difficult.

Future plans

In PULS, one established direction that the group is exploring is the acquisition of different types of knowledge bases in parallel, where they provide mutual constraints to maintain high precision. An important new direction is extending the linguistic coverage of PULS; most of the work in the field still focuses on English-language text, motivated by the abundance of lower-level processing tools available for English. PULS has been working to extend the coverage to Russian, which is notably much more poor in NLP resources, yet has high value in the news scenarios covered to date. The methods employed include bootstrapping techniques, together with methods for "projecting" existing (English) resources/tools onto the target languages, and using similar documents in English and Russian, so-called "comparable" data (i.e., not strictly "parallel" data).

In Etymon, the methods we devise for examining relationships are being applied in a wider context:

a. extending alignment of observed data to performing reconstruction of the corresponding forms in the unobserved, ancestral languages.
b. to apply the methodology to other language families, especially the less-studied families, to help obtain novel results.
c. collaboration with population-genetics experts, to i. explore cross-pollination on the methodological level (i.e., applicability of algorithms to different kinds of data), and ii. explore complementary information carried in linguistic vs. genetic data. This is available (to a degree) for certain population, and presents an exciting opportunity for deeper insight into the inter-relationships among the related groups.

Societal, economic and technical impact

In the domain of Epidemic Surveillance, PULS collaborates with major National and international Health Agencies, including Health Ministries of several European Countries (France, UK, and Spain), the European Center for Disease Control (ECDC), in Stockholm, Sweden, and the WHO (World Health Organisation). Computational methods developed in PULS help specialists at these agencies perform their tasks more efficiently, in protecting the European citizen from dangerous epidemics. Users of PULS in the domain of business intelligence include international companies.

The results of Etymon provide insights into the structure and development of language families. These relationships have been studied for two centuries using manual methods. Computational methods for analysing these fascinating data are only now beginning to emerge. The results will have wide-ranging implications for the understanding of the common origins of language.

In FinUgRevita, we work with interested parties, teachers and schools, in areas where there is interest in supporting or revitalising languages in danger of imminent disappearance, which has the potential to slow or reverse language decay.

Cooperation

Cooperation within Algodan

Link and Pattern Discovery Group: PULS exploits the tools developed by the Biomine group for analysing the complex graphs resulting from the analysis of news in the domain of business intelligence.

Cooperation within the University

Etymon project collaborates with the Population Genetics group, lead by Prof. Jukka Corander, (joint publications currently in preparation), as well with researchers at the Department of Modern Languages, and the Department of Finno-Ugric Studies.

International cooperation

PULS and Etymon have very strong on-going international collaboration.

In PULS we have been collaborating with the Text Mining Research Unit at the European Commission's Joint Research Centre (JRC, in Ispra, Italy) for several years. In the domain of border security, PULS collaborates with the EC Frontex Agency for the protection of the European External Frontiers. PULS provides an on-line feed into MedISys, the system for global news surveillance developed by the JRC. Our results and links to our databases are served in real-time on the MedISys platform at http://medusa.jrc.it – JRC sends raw articles that it mines from the Web to PULS, and PULS returns the results to JRC. MedIsys has thousands of users every day, government and private; our results are documented in several joint publications. Some of these users – Public Health professionals – use PULS on a daily basis. PULS participates in the Global Health Security Initiative (GHSI)/Global Health Security Action Group's Early Alerting and Reporting project. The GHSI is an international consortium created to strengthen health preparedness and response globally to biological, chemical, radiological/nuclear threats. This initiative was launched in 2001 by Canada, the EU, France, Germany, Italy, Japan, Mexico, the UK, and the USA. PULS is the only European academic partner within the GHSAG consortium (the other two are from University of Tokyo and Harvard). Other members in the consortium are user organisations – those who need high-quality analysis of news for surveillance of thousands of on-line sources, in real time. The partners include the European Center for Disease Control (ECDC) in Stockholm, Sweden, the WHO (World Health Organisation), as well as several large national health ministries. As mentioned, these users are important for our work for providing challenging scenarios and good data.

Etymon is an inter-disciplinary project, where in addition to local collaborators; we work closely with the Russian Academy of Sciences, in Moscow, Russia.

Selected publications

1. DM Hartley, NP Nelson, RR Arthur, P Barboza, N Collier, N Lightfoot, JP Linge, E van der Goot, A Mawudeku, LC Madoff, L Vaillant, R Walters, R Yangarber, J Mantero, CD Corley, JS Brownstein. Overview of Internet biosurveillance. In Journal of Clinical Microbiology and Infection, 19(6) (2013).
2. Hannes Wettig, Javad Nouri, Kirill Reshetnikov, Roman Yangarber. Information-theoretic modeling of etymological sound change. In Approaches to measuring linguistic differences (Lars Borin, Anju Saxena, eds.) (2013) Trends in Linguistics Series, Volume 265. Mouton de Gruyter.
3. Javad Nouri, Lidia Pivovarova, Roman Yangarber. MDL-based models for transliteration generation. In SLSP-2013: International Conference on Statistical Language and Speech Processing. Springer Verlag, Lecture Notes in Artificial Intelligence (LNAI) Volume 7978. (2013) Tarragona, Spain.

4. Silja Huttunen, Arto Vihavainen, Mian Du, Roman Yangarber. Predicting Relevance of Event Extraction for the End User. In "Multi-source, Multilingual Information Extraction and Summarization", Theory and Applications of Natural Language Processing, T. Poibeau et al. (eds.). Springer-Verlag (2012) Berlin, Heidelberg.
5. Lidia Pivovarova, Mian Du, Roman Yangarber.Adapting the PULS event extraction framework to analyze Russian text. In ACL/BSNLP-2013: 4th Biennial Workshop on Balto-Slavic Natural Language Processing. (2013) Sofia, Bulgaria.
6. Hannes Wettig, Suvi Hiltunen, Roman Yangarber. MDL-based modeling of etymological sound change in the Uralic language family. WITMSE-2011: The Fourth Workshop on Information Theoretic Methods in Science and Engineering (2011) Helsinki, Finland.
7. Martin Atkinson, Jakub Piskorski, Erik Van der Goot, Roman Yangarber. Multilingual real-time event extraction for border security intelligence gathering. Counterterrorism and Open Source Intelligence (Lecture Notes in Social Networks, Vol. 2. Uffe Kock Wiil, editor). Springer. pp. 355-390 (2011).
8. Hannes Wettig, Suvi Hiltunen, Roman Yangarber. MDL-based models for aligning etymological data. RANLP-2011: Conference on Recent Advances in Natural Language Processing (2011) Hissar, Bulgaria.
9. Silja Huttunen, Arto Vihavainen, Peter von Etter, Roman Yangarber. Relevance prediction in information extraction using discourse and lexical features. Nodalida-2011: Nordic Conference on Computational Linguistics (2011) Riga, Latvia.

Team Link and Pattern Discovery

Discovery Group: Data Mining and Computational Creativity

Members

- Hannu Toivonen, Professor, Group leader
- Alessandro Valitutti, PhD, Postdoctoral researcher (-2013)
- Laura Langohr, Doctoral student
- Oskar Gross, Doctoral student
- Jukka Toivanen, Doctoral student
- Fang Zhou, Doctoral student (-2012)
- Esther Galbrun, Doctoral student, co-supervised with Mikko Koivisto (-2013)
- External doctoral students
    - Joonas Paalasmaa, Doctoral student, employed by Beddit.com Ltd. (-2014)
    - Mika Timonen, Doctoral student, employed by VTT, the Technical Research Centre of Finland (-2013)

Mission of the group

The Discovery group develops novel methods and tools for data mining and computational creativity. Our focus is on algorithmic methods for discovering links and patterns in data, and especially on their use in creative systems. Application areas range from computational generation of poetry to link discovery in bioinformatics and to sleep analysis.

Research activities

A methodological focus area has been in analysis and exploration methods for weighted (biological or word co-occurrence) graphs, i.e., mainly in the theme "L - learning from and mining structured and heterogeneous data" of Algodan.

In 2011, we initiated work on computational poetry and have since expanded our efforts to computational humor, music, and fine arts [3, 4, 7]. Especially in linguistic creativity we develop novel methods that learn from existing texts and thereby minimize the need for manually coded or language specific knowledge, making the methods more easily adabtable to different languages. Artistic results of the computational creativity work have been published in print and exhibited in several galleries (see the section on Societal, economical and technical impact below).

Various (other) data mining method development activities have also been continued, covering graph mining [12] and its applications in biology [1, 9, 11], sleep analysis [7, 8], discovery of functional dependencies [5], subgroup discovery [6], redescription mining [10], and news analysis.

Future plans

The group will focus on computational creativity, using data mining and graph mining as the methodological basis. We will primarily develop methods that either are “knowledge poor” (in the sense of not relying on knowledge bases or manually coded knowledge) or that help make sense of data [7].

In the past two years we have established strong networks with computational creativity researchers internationally (see below). In the future, we plan to establish contacts and collaboration with scientists in applied fields (literature, cinema), also domestically.

Societal, economical and technical impact

With our shift of research focus towards computational creativity, our work has reached two novel forms of impact.

First, we have a societal and cultural impact via contributions to *artistic acts* using computational creativity. Since 2012, we have contributed computational poetry to an art book, given several performances of biomusic, participated in installations of brain wave inspired poetry and of biophysical sensor based music, and contributed to paintings, among others. Most of these works have been collaborations with professional artists and have been exhibited or performed in galleries. (For more information, see http://www.cs.helsinki.fi/en/discovery/art .)

Second, the impact of this work is indirectly evidenced by the attention received from public media. Since 2012, our work has been covered on TV (YLE, Finland), radio (CBC, Canada), popular press (e.g., The Times, UK; New Scientist, USA; Helsingin Sanomat, Finland), leading technology web sites (e.g., CNET, Engadget), and hundreds of other websites in at least fifteen languages.

Economical and technical impact is primarily via collaboration with companies (see above). Sleep research carried out with Beddit Ltd. seems highly promising based on the interest to their newest product.

Cooperation

Cooperation within Algodan

Co-supervision of students, exchange of information, and other informal co-operation.

Cooperation within the universities

- Prof. Markku Partinen, Institute of Clinical Medicine: joint research in sleep analysis, joint publications.
- Dr. Juhani Huovelin, Department of Physics: joint research on news analysis and aggregation, joint publications.
- Dr. Liisa Ilomäki, Institute of Behavioural Sciences: joint proposal to research educational uses of computational creativity in schools.
- Dr. Alina Leminen, Institute of Behavioural Sciences: starting joint research on empirical cognitive neuroscience studies for linguistic creativity.

Other national cooperation

- Beddit Ltd: research collaboration on sleep analysis methods, supervision of PhD studies, joint publications.
- Research collaboration with a number of companies in the Future Media program of the ICT cluster of the Finnish Strategic Centres for Science, Technology and Innovation (Tivit Ltd).

International cooperation

- Concept Creation Technology, ConCreTe, EU project with 7 partners, 2013-16.
- Promoting the Scientific Exploration of Computational Creativity, PROSECCO, EU coordination action with 7 partners, 2013-16.
- Professor Nada Lavrac, Jozef Stefan Institute, Slovenia: joint research on data and text mining for bioinformatics. Joint publications.
- Professor Ross D. King, University of Machester, UK: joint research on graph mining in bioinformatics, research visits, joint publication.

- Professor Jiuyong Li, University of Southern Australia: joint research on dependency mining. Joint publications, research visits.
- Assoc. Professor Antoine Doucet, University of Caen, France: joint research on text mining, student exchange, research visits, joint publications.

Selected publications

1. The Use of Weighted Graphs for Large-Scale Genome Analysis. Fang Zhou, Hannu Toivonen, Ross D. King. PLoS One, accepted for publication.
2. Software Newsroom – an approach to automation of news search and editing. Juhani Huovelin, Oskar Gross, Otto Solin, Krister Lindén, Sami Maisala, Tero Oittinen, Hannu Toivonen, Jyrki Niemi, and Miikka Silfverberg. Journal of Print Media Technology Research, in press.
3. "Let Everything Turn Well in Your Wife": Generation of Adult Humor Using Lexical Constraints. Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, Jukka M. Toivanen. The 51st Annual Meeting of the Association for Computational Linguistics (ACL), Sofia, Bulgaria, August 2013.
4. Harnessing Constraint Programming for Poetry Composition. Jukka Toivanen, Matti Järvisalo, Hannu Toivonen. The Fourth International Conference on Computational Creativity (ICCC), Sydney, Australia, June 2013.
5. Effective Pruning for the Discovery of Conditional Functional Dependencies. Jiuyong Li, Jiuxue Liu, Hannu Toivonen, Jianming Yong. The Computer Journal 56 (3): 378-392. 2013.
6. Contrasting Subgroup Discovery. Laura Langohr, Vid Podpecan, Marko Petek, Igor Mozetic, Kristina Gruden, Nada Lavrac, Hannu Toivonen. The Computer Journal 56 (3): 289-303. 2013.
7. Sleep Musicalization: Automatic Music Composition from Sleep Measurements. Aurora Tulilaulu, Joonas Paalasmaa, Mikko Waris, Hannu Toivonen. Eleventh International Symposium on Intelligent Data Analysis (IDA), LNCS 7619, 392-403, Helsinki, Finland, October 2012. (See sleepmusicalization.net for an implementation and example music.) (Winner of the IDA 2012 Frontier Prize as the 'most novel and visionary contribution' of the conference.)
8. Unobtrusive Online Monitoring of Sleep at Home. Joonas Paalasmaa, Mikko Waris, Hannu Toivonen, Lasse Leppäkorpi, Markku Partinen. International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), 3784-3788, San Diego, USA, August-September 2012.
9. Biomine: Predicting links between biological entities using network models of heterogeneous databases, Lauri MA Eronen and Hannu TT Toivonen. BMC Bioinformatics 13:119, 2012. (Indicated as 'Highly accessed' by the journal.)
10. From Black and White to Full Colour: Extending Redescription Mining Outside the Boolean World. Esther Galbrun, Pauli Miettinen. In SIAM International Conference on Data Mining (SDM), 2011. Extended version to appear in Statistical Analysis and Data Mining.
11. SegMine workflows for semantic microarray data analysis in Orange4WS, Vid Podpecan, Nada Lavrac, Igor Mozetic, Petra Kralj Novak, Igor Trajkovski, Laura Langohr, Kimmo Kulovesi, Hannu Toivonen, Marko Petek, Helena Motaln, Kristina Gruden. BMC Bioinformatics 12:416, 2011.
12. Compression of weighted graphs, Hannu Toivonen, Fang Zhou, Aleksi Hartikainen, Atte Hinkka. In The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), San Diego, CA, USA, August 2011.

Team Machine Learning

Machine Learning Group

Machine Learning group consists of two subgroups:

- Kernel Machines, Pattern Analysis and Computati onal Biology (until 2011: Computational Systems Biology and Bioinformatics), and
- Learning Theory.

Members

- Jyrki Kivinen, Professor, Principal Investigator, Group leader (subgroup Learning Theory)
- Juho Rousu, Professor, Principal Investigator, Group leader (subgroup Kernel Machines, Pattern Analysis and Computational Biology)
- Jana Kludas, Postdoctoral researcher (2012-)
- Esa Pitkänen, Postdoctoral researcher (-2011)
- Markus Heinonen, Doctoral student (-2012)
- Panu Luosto, Doctoral student (-2013)
- Hongyu Su, Doctoral student (2011-)
- Huibin Shen, Doctoral student (2012-)
- Anna Cichonska, Doctoral student (2013-)

Mission of the group

The group develops machine learning methods, models and tools for computational sciences, in particular computational biology. The methodological backbone of the group is kernel methods and regularized learning. The group particularly focusses in learning with multiple and structured targets, multiple views and ensembles. Applications of interest in computational biology include network reconstruction, gene functional classification as well as metabolite identification.

Research activities

L - Learning from and mining structured and heterogeneous data

Metabolite identification from tandem mass spectra is an important problem in metabolomics, underpinning metabolic modelling and network analysis. With the support of Academy of Finland research grant (MIDAS, 2013-2017), we develop methods for automatic identification of metabolites from tandem mass spectra using machine learning. In our most recent approach, fragmentation trees are computed from the spectra, and turned into a collection of kernels capturing similarities of trees from different viewpoints. The kernels are then combined using state-of-the-art multiple kernel learning methods. Our method is currently the most accurate fully computational approach for metabolite identification [1].

In metabolic network analysis we developed methods for simultaneous reconstruction of metabolic networks for a set of related organism, connected by a phylogenetic tree. The method generalizes the Fitch-Hartigan algorithm to discover phylogenetic tree with minimum number of reaction mutations so that each ancestral node corresponds to a gapless metabolic network [2]. Another example of our recent work is biomarker discovery from plasma proteomics and clinical data using sparse canonical correlation analysis. In the method L1-penalization is used for the proteomics data while the clinical data is kernelized [4].

F - Foundations of algorithmic data analysis

In machine learning, we have developed new methods for multilabel classification, relying on ensemble learning on a collection of random output graphs imposed on the multilabels, and a kernel-based structured output learner as the base classifier. Theoretically, we have been able to show that our multilabel ensemble not only benefits from diversity in the base classifier predictions (phenomenon already known for single target ensembles), but also from the covariance of predictions of microlabel pairs. Experimentally, our method has robust performance on a variety tasks, matching or exceeding the performance of commonly used multilabel/task learning approaches [3].

Issues related to the Minimum Description Length (MDL) principle have been studies in joint work between Panu Luosto and Dr. Petri Kontkanen. The main conceptual novelty is a systematic way of dealing with cases where the parametric complexity of the model class is infinite. The general principle has been specifically applied mainly in clustering and histogram models. There has also been some work on more specific computational issues that arise in these applications [5].

Future plans

Major themes for future research of the group include:

- Kernel methods for structured data. We will develop kernel representations for structured objects such as sequences, trees and graphs, and efficient algorithms for mapping data into kernels and back (aka pre-image problem). Especially we focus on predicting structured output, a setup that aims to leverage the structure of the data to increase efficacy of learning and prediction. Learning with multiple kernels and ensembles is another direction of importance.

- Network labeling. In many applications, the data to be predicted has a network form. In particular, network labeling problems involve a known network structure, and a set of data instances that each activates a set of nodes. The prediction task is to learn which nodes are to be activated for a given input data. Examples of such prediction problems are frequent in document management (e.g. hierarchical text classification) as well as information and computer networks (e.g. resource placement).

- Computational biology. A major application field of our methods is in biological sciences. Our core competence there is in biological network reconstruction and analysis, especially in metagenomic context, as well as prediction problems involving small molecules (e.g. metabolomics, mass spectrometry).

Social, economical and technical impact

Our research is geared towards building computational tools and methods for the analysis of data arising in biotechnology and biomedicine. The impact of our work is in making applied research better targeted, faster and more cost-effective. An example of this is our metabolite identification technology, which speeds up research in metabolomics by using less time in verifying false positive predictions and may have ramifications to applications such as doping control and forensics.

Cooperation

Cooperation within Algodan

The group collaborates with the Succinct Data Structures (currently: Genome Scale Algorithmics) group in algorithms and index structures for graph data and data mining group in network labeling problems.

National cooperation

- University of Helsinki, Finland/Professor Liisa Holm. Collaboration in enzyme function prediction and metabolic reconstruction [2].
- VTT Technical Research Center of Finland/Professor Merja Penttilä, Doc. Merja Itävaara. Collaboration in pathway modelling in industrial microbes (EU FP7 STREP BIOLEDGE) and deep biosphere microbiota (GEOBIOINFO project, 2011-).
- University of Helsinki (Complex Systems Computation Group). Ongoing collaboration on MDL with Dr. Petri Kontkanen, which has lead to three published articles.

International cooperation

- University College London, United Kingdom/Professor John Shawe-Taylor. Collaboration in kernel methods and structured output learning. The collaboration has resulted in several new methods for machine learning for structured data, including sequences, taxonomies, and general graphs.
- Friedrich-Schiller Universität Jena/Professor Sebastian Böcker. Collaboration in Metabolite identification method development.
- ETH Zurich, Institue of Molecular Systems Biology/Dr. Nicola Zamboni. Collaboration in Mass Spectrometric data analysis

Selected publications

1. Shen, H., Dührkop, K., Böcker, S., Rousu, J. Metabolite Identification through Multiple Kernel Learning on Fragmentation Trees. 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2014), accepted.
2. Pitkänen, E., Jouhten, P., Hou, J., Fahad, M.S., Blomberg, P., Kludas, J., Oja, M., Holm, L., Penttilä, M., Rousu, J., Arvas, M. (2014). Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species. PLoS Computational Biology 10, 2, p. e1003465
3. Su, H., Rousu, J. (2013) Multilabel Classification through Random Graph Ensembles. 5th Asian Conference on Machine Learning, JMLR W&CP 29, pp. 404-418
4. Rousu, J., Agranoff, D., Sodeinde, O., Shawe-Taylor, J., & Fernandez-Reyes, D. (2013). Biomarker Discovery by Sparse Canonical Correlation Analysis of Complex Clinical Phenotypes of Tuberculosis and Malaria. PLoS Computational Biology 9, 4, p. e1003018
5. Luosto, P. (2013) Normalized maximum likelihood methods for clustering and density estimation. PhD thesis, Report A-2013-8, Department of Computer Science, University of Helsinki

Team Neuroinformatics

Neuroinformatics Group

Members

- Aapo Hyvärinen, Professor, Group leader
- Patrik Hoyer, Academy Research Fellow, Co-leader (-7/2013)
- Michael Gutmann, Postdoctoral researcher (-5/2012)
- Jun-ichiro Hirayama, Postdoctoral researcher (-6/2011)
- Hiroaki Sasaki, Postdoctoral researcher (10/2011-9/2012)
- Cristina Campi, Postdoctoral researcher (-8/2011)
- Hugo Eyeherabide, Postdoctoral researcher (10/2011-)
- Jukka-Pekka Kauppi, Postdoctoral researcher (6/2011-)
- Doris Entner, Doctoral student (-11/2013)
- Antti Hyttinen, Doctoral student (-5/2013)
- Jouni Puuronen, Doctoral student
- Miika Pihlaja, MSc/Doctoral student (-12/2011)

Mission of the group

Our mission is to develop statistical data analysis methods, with the particular applications of neuroscience in mind. In some areas of neuroscience, such as brain imaging, measurement devices provide huge amounts of data and new methods are needed to analyze the data. On the other hand, modelling perception can be approached from a Bayesian viewpoint, as probabilistic modelling of typical stimuli. General-purpose statistical learning methods are naturally developed at the same time.

Research activities

D - Discovery of hidden structure in high-dimensional data

Testing for ICA. In the ICA research community, the almost exclusive focus has been on estimation, and testing of the results has received almost zero attention. However, in any practical application it would be extremely important to be able to test some kind of statistical significance of the components. Otherwise, we don't know if the results are just due to random noise, or local minima. We have developed a new testing framework for ICA based on the idea of having many datasets (or splitting one dataset into many), applying ICA separately on each dataset, and then investigating if the results are similar enough in the different datasets. We were able to formulate a proper null hypothesis and use the conventional machinery of the theory of statistical hypothesis testing. The test was developed for two different cases: testing the mixing matrix [1], or testing the actual values of the independent components [2].

Causal discovery. Our project on estimating causal relations from continuous-valued data continued to be successful and lead to many new methods. Highlights include: very simple measures of causal direction for two variables, i.e. does x cause y or does y cause x [3], a procedure to estimate the strength of causal effects in the presence of hidden variables [4], and identifiability results of linear cyclic causal models based on randomized experiments [5].

Neuroimaging data analysis. In our joint neuroimaging project with Prof. Riitta Hari of Aalto University, we investigated new variants of ICA for analysis of spontaneous brain activity (e.g. during rest) [6], as well as "decoding" in MEG, i.e. using classification methods to infer what kind of stimulation was given to the

brain, using only the MEG data as input to the classifier [7]. An important focus was the very deep topic of *two-person neuroscience*. This project, related to Prof. Hari's ERC Advanced Grant of the same title, attempts to open completely new vistas in social neuroscience by measuring two subjects' brain activities at the same time. This is an extremely challenging, high-risk project since almost everything has to be built from scratch: instrumentation (connecting to MEG systems with audio and video links), experimental design (since such experiments have hardly been conducted before), and data analysis (which is our responsibility in this joint project). Initial developments in the machine learning methods were published in [8], but much remains to be done.

F - Foundations of algorithmic data analysis

Estimation of non-normalized probabilistic models. Our project on estimating non-normalized probabilistic models culminated in a long JMLR paper [9] which shows a deep connection between supervised and unsupervised learning, and how it can be utilized to estimate such intractable models.

Future plans

The end of the Algodan funding period coincides with a changepoint in the structure and focus of the neuroinformatics group, since the co-leader, Patrik Hoyer, left academia to start his own company in autumn 2013. At the same time, the long sabbatical that Aapo Hyvärinen spent in ATR, Japan, in 2013-2014 led to many new ideas. In the future, causal discovery and estimation of non-normalized probabilistic models will be more or less discontinued, and replaced by new projects related to modeling spontaneous brain activity. Measuring spontaneous brain activity (i.e. during rest) is now popular in brain imaging, but functional models of the information processing happening in spontaneous brain activity are very rare. In other words, we don't really know very well what the brain is doing during rest, and, most importantly, why. Theoretical models, based on machine learning and probabilistic modeling, would be extremely useful here.

Societal, economical and technical impact

Several of the new methods have been made publicly available as software packages distributed on the internet.

Cooperation

Cooperation within Algodan

In recent years, one of the main strands of work in the group has been on the topic of methods for learning directed graphical models from data. This family of models includes Bayesian networks, well studied in the fields of machine learning and artificial intelligence. Thus, the group has recently benefited greatly from the expertise of Dr. Mikko Koivisto on exact algorithms for structure learning of Bayesian networks. Recently, we have also benefited from discussions with Dr. Koivisto and Dr. Petteri Kaski on problems in combinatorics derived from our identifiability conditions on learning cyclic models from randomized experiments.

National cooperation

- Riitta Hari, Aalto University, Finland. Topic: Brain imaging. Joint project funded by Academy of Finland; joint Doctoral student, joint Postdoctoral researchers

International cooperation

- Stephen M. Smith, Oxford University, UK. Topic: causal discovery in fMRI. Aapo Hyvärinen visited him for two months in 2011. One joint publication.

- Frederick Eberhardt, Peter Spirtes, et al, Carnegie Mellon University, USA. Topic: theory of causal discovery. Several visits (2-3 weeks each) by Patrik Hoyer. Several joint publications.
- Dominik Janzing, Max Planck Institute for Cybernetics, Germany. Topic: causal inference among high-dimensional variables. Joint publication.
- Alessio Moneta and Alex Coad, Max Planck Institute for Economics, Germany. Topic: causal inference in time-series. Joint publications.
- Shin Ishii, Kyoto University and Motoaki Kawanabe, ATR, Japan. Topic: Brain imaging signal analysis. Aapo Hyvärinen visited ATR for 10 months in 2013-2014. We have a joint publication and a couple in preparation.
- Klaus-Robert Müller, Berlin Institute of Technology, Germany. Topic: Decoding biomedical signals. Jukka-Pekka Kauppi spent 2 months in Berlin in 2013 to start the collaboration. One joint article in preparation.
- Masashi Sugiyama, Tokyo Institute of Technology, Japan. Topic: Image processing. We have a joint project based on joint post-doc supervision. One joint article submitted.

Selected publications

1. A. Hyvärinen. Testing the ICA mixing matrix based on inter-subject or inter-session consistency. NeuroImage, 58(1):122-136, 2011.
2. A. Hyvärinen and P. Ramkumar. Testing independent component patterns by inter-subject or inter-session consistency, Frontiers in Human Neuroscience, 7:94, 2013.
3. A. Hyvärinen and S. M. Smith. Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. J. of Machine Learning Research, 14:111-152, 2013.
4. A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. J. of Machine Learning Research, 13:3387–3439, 2012.
5. A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Experiment Selection for Causal Discovery. J. of Machine Learning Research, 14:3041–3071, 2013.
6. P. Ramkumar, L. Parkkonen, R. Hari, and A. Hyvärinen. Characterization of neuromagnetic brain rhythms over time scales of minutes using spatial independent component analysis. Human Brain Mapping, 33(7):1648-1662, 2012.
7. J.-P. Kauppi, L. Parkkonen, R. Hari, and A. Hyvärinen. Decoding MEG rhythmic activity using spectrospatial information. NeuroImage, 83:921-936, 2013.
8. C. Campi and L. Parkkonen and R. Hari and A. Hyvärinen. Non-linear canonical correlation for joint analysis of MEG signals from two subjects. Frontiers in Brain Imaging Methods 7:107, 2013.
9. M. U. Gutmann and A. Hyvärinen. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics, J. Machine Learning Research 13:307-361, 2012.

5. Publications

2013

Articles in refereed scientific journals

1. P.R. Adhikari and J. Hollmén. Fast progressive training of mixture models for model selection. *Journal of Intelligent Information Systems*, 2013, pp. 1-19.
2. P. Barboza, L. Vaillant, A. Mawudeku, N.P. Nelson, D.M. Hartley, L.C. Madoff, J.P. Linge, N. Collier, J.S. Brownstein, R. Yangarber, P. Astagneau. Evaluation of Epidemic Intelligence Systems Integrated in the Early Alerting and Reporting Project for the Detection of A/H5N1 Influenza Events. *PLoS One*, vol. 8, no. 3, 2013, pp. e57252.
3. M. Bonke, M. Turunen, M. Sokolova, A. Vaharautio, T. Kivioja, M. Taipale, M. Björklund and J. Taipale. Transcriptional Networks Controlling the Cell Cycle. *G3 - Genes genomes genetics*, vol. 3, no. 1, 2013, pp. 75-90.
4. Y. Borodin, V. Polishchuk, J. Mahmud, I.V. Ramakrishnan and A. Stent. Live and learn from mistakes: A lightweight system for document classification. *Information Processing & Management*, vol. 49, no. 1, 2013, pp. 83-98.
5. R.P.J.C. Bose, W.M.P. van der Aalst, I. Zliobaite and M. Pechenizkiy. Dealing with Concept Drifts in Process Mining. *IEEE Transactions on Neural Networks and Learning Systems*, 25(1), 2013, pp. 154-171.
6. A.G. Bunn, E. Jansma, M. Korpela, R.D. Westfall and J. Baldwin. Using simulations and data to evaluate mean sensitivity ($$) as a useful statistic in dendrochronology. *Dendrochronologia*, 31(3), 2013, pp. 250-254.
7. C. Campi, L. Parkkonen, R. Hari and A. Hyvärinen. Non-linear canonical correlation for joint analysis of MEG signals from two subjects. *Frontiers in Brain Imaging Methods*, vol. 7, no. 107, 2013.
8. H.G. Eyherabide and I. Samengo. When and why noise correlations are important in neural decoding. *Journal of Neuroscience*, vol. 33, no. 45, 2013, pp. 17921-17936.
9. T. Gagie, J. Kärkkäinen, G. Navarro and S.J. Puglisi. Colored range queries and document retrieval. *Theoretical Computer Science*, vol. 483, 2013, pp. 36-50.
10. E. Giaquinta and S. Grabowski. New algorithms for binary jumbled pattern matching. *Information Processing Letters*, vol. 113, no. 14-16, 2013, pp. 538-542.
11. E. Giaquinta, S. Grabowski and K. Fredriksson. Approximate pattern matching with k-mismatches in packed text. *Information Processing Letters*, vol. 113, no. 19-21, 2013, pp. 693-697.
12. E. Giaquinta, S. Grabowski and E. Ukkonen. Fast Matching of Transcription Factor Motifs Using Generalized Position Weight Matrix Models. *Journal of Computational Biology*, vol. 20, no. 9, 2013, pp. 621-630.
13. E. Giaquinta and L. Pozzi. An Effective Exact Algorithm and a New Upper Bound for the Number of Contacts in the Hydrophobic-Polar Two-Dimensional Lattice Model. *Journal of Computational Biology*, vol. 20, no. 8, 1013, pp. 593-609.
14. M.U. Gutmann and A. Hyvärinen. A three-layer model of natural image statistics. *Journal of Physiology (Paris)*, vol. 107, no. 5, 2013, pp. 369-398.
15. J. Huovelin, O. Gross, O. Solin, K. Linden, S.P.T. Maisala, T. Oittinen, H. Toivonen, J. Niemi and M. Silfverberg. Software Newsroom – an approach to automation of news search and editing. *Journal of Print Media Technology research*, vol. 2, no. 3, 2013, pp. 141-156.
16. H. Huttunen, T. Manninen, J.-P. Kauppi and J. Tohka. Mind reading with regularized multinomial logistic regression. *Machine Vision & Applications*, vol. 24, no. 6, 2013, pp. 1311-1325.
17. A. Hyttinen, F. Eberhardt and P.O. Hoyer. Experiment Selection for Causal Discovery. *Journal of Machine Learning Research*, vol. 2013, no. 14, 2013, pp. 3041-3071.
18. A. Hyvärinen and P. Ramkumar. Testing independent component patterns by inter-subject or inter-session consistency. *Frontiers in Human Neuroscience*, vol. 7, 2013.

19. A. Hyvärinen and S.M. Smith. Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. *Journal of Machine Learning Research*, vol. 14, 2013, pp. 111-152.
20. A. Jolma, J. Yan, T. Whitington, J. Toivonen, K.R. Nitta, P. Rastas, E. Morgunova, M. Enge, M.J.O. Taipale, G. Wei, K. Palin, J.M. Vaquerizas, R. Vincentelli, N.M. Luscombe, T.R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja and J. Taipale. DNA-Binding Specificities of Human Transcription Factors. *Cell*, vol. 152, no. 1-2, 2013, pp. 327-339.
21. J.-P. Kauppi, L. Parkkonen, R. Hari and A. Hyvärinen. Decoding magnetoencephalographic rhythmic activity using spectrospatial information. *NeuroImage*, vol. 83, 2013, pp. 921-936.
22. L. Langohr, V. Podpečan, M. Petek, I. Mozetič, K. Gruden, N. Lavrač and H. Toivonen. Contrasting Subgroup Discovery. *Computer Journal*, vol. 56, no. 3, 2013, pp. 289-303.
23. J. Li, J. Liu, H. Toivonen and J. Yong. Effective Pruning for the Discovery of Conditional Functional Dependencies' *Computer Journal*, vol. 56, no. 3, 2013, pp. 378-392.
24. J.M. Leppilahti, M.A. Kallio, T. Tervahartiala, T. Sorsa and P. Mantyla. Gingival Crevicular Fluid (GCF) Matrix Metalloproteinase-8 Levels Predict Treatment Outcome Among Smoking Chronic Periodontitis Patients. *Journal of Periodontology*, 85, 2013.
25. A. Moneta, D. Entner, P.O. Hoyer and A. Coad. Causal Inference by Independent Component Analysis: Theory and Applications. *Oxford Bulletin of Economics and Statistics*, vol. 75, no. 5, 2013, pp. 705-730.
26. E. Packer, P. Bak, M. Nikkilä, V. Polishchuk and H.-J. Ship. Visual Analytics for Spatial Clustering: Using a Heuristic Approach for Guided Exploration. *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, 2013, pp. 2179-2188.
27. P. Parviainen and M. Koivisto. Finding Optimal Bayesian Networks Using Precedence Constraints. *Journal of Machine Learning Research*, vol. 14, 2013, pp. 1387-1415.
28. P. Rastas, L. Paulin, I. Hanski, R.J. Lehtonen and P. Auvinen. Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics*, vol. 29, no. 24, 2013, pp. 3128-3134.
29. J. Rousu, D.D. Agranoff, O. Sodeinde, J. Shawe-Taylor and D. Fernandez-Reyes. Biomarker Discovery by Sparse Canonical Correlation Analysis of Complex Clinical Phenotypes of Tuberculosis and Malaria. *PLOS Computational Biology*, 9(4), 2013, 1003018.
30. H. Sasaki, M.U. Gutmann, H. Shouno and A. Hyvärinen. Correlated topographic analysis: estimating an ordering of correlated components' *Machine Learning*, vol. 92, no. 2-3, 2013, pp. 285-317.
31. H. Shen, N. Zamboni, M. Heinonen and J. Rousu. Metabolite Identification through Machine Learning — Tackling CASMI Challenge Using FingerID. *Metabolites*, 3(2), 484-505.
32. J. Yan, M. Enge, T. Whitington, K. Dave, J. Liu, I. Sur, B. Schmierer, A. Jolma, T. Kivioja, M. Taipale and J. Taipale. Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell*, vol. 154, no. 4, 2013, pp. 801-813.

Refereed conference articles and articles in edited books

1. P.R. Adhikari and J. Hollmén. Mixture Models from Multiresolution 0-1 Data. In J. Fürnkranz, E. Üllermeier and T. Higuchi (editors), *Proceedings of Sixteenth International Conference on Discovery Science (DS 2013)*, 2013, Springer Berlin Heidelberg, pp. 1-16.
2. M. Atkinson, M. Du, J. Piskorski, H. Tanev, R. Yangarber and V. Zavarella. Techniques for Multilingual Security-related Event Extraction from Online News. In A. Przepiórkowski et al.(editors), *Computational Linguistics—Applications, Studies in Computational Intelligence*, vol. 458, 2013, Springer-Verlag, pp. 163-186.
3. P. Austrin, P. Kaski, M. Koivisto and J. Määttä. Space–Time Tradeoffs for Subset Sum: An Improved Worst Case Algorithm. In *Proceedings of the 40th International Colloquium on Automata, Languages, and Programming (ICALP 2013), Part I*, Riga, Latvia, July, 2013, Lecture Notes in Computer Science, vol. 7965, pp. 45-56.
4. D. Belazzougui, F. Cunial, J. Kärkkäinen and V. Mäkinen. Versatile succinct representations of the bidirectional Burrows-Wheeler transform. In *Proceedings of the 21st Annual European Symposium*

on Algorithms (ESA 2013), Sophia Antipolis, France, September 2-4, 2013, Lecture Notes in Computer Science, vol. 8125, pp. 133-144.
5. D. Belazzougui, T. Gagie and G. Navarro. Better Space Bounds for Parameterized Range Majority and Minority. In *Proceedings* of *13th International Symposium on Algorithms and Data Structures (WADS 2013)*, London, ON, Canada, August 12-14, 2013, Lecture Notes in Computer Science, vol. 8037, pp. 121-132.
6. A. Bifet, J. Read, I. Zliobaite, B. Pfahringer and G. Holmes. Pitfalls in benchmarking data stream classification and how to avoid them. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2013, pp. 465-479.
7. A. Bifet, J. Read, B. Pfahringer, G. Holmes and I. Zliobaite. CD-MOA: Change Detection Framework for Massive Online Analysis. In *Proceedings of the 20th International Symposium on Intelligent Data Analysis (IDA 2013)*, 2013, pp. 92-103.
8. F. Cicalese, T. Gagie, E. Giaquinta, E.S. Laber, Z. Lipták, R. Rizzi and A. Tomescu. Indexes for Jumbled Pattern Matching in Strings, Trees and Graphs. In *Proceedings of the 20th International Symposium on String Processing and Information Retrieval (SPIRE 2013)*, Jerusalem, Israel, October 7-9, 2013, Lecture Notes in Computer Science, vol. 8214, pp. 56-63.
9. M. Crochemore, R. Grossi, J. Kärkkäinen and G.M. Landau. A Constant-Space Comparison-Based Algorithm for Computing the Burrows-Wheeler Transform. In *Proceedings of the 24th Annual Symposium Combinatorial Pattern Matching (CPM 2013*), Bad Herrenalb, Germany, June 17-19, 2013, Lecture Notes in Computer Science, vol. 7922, pp. 74-82.
10. D. Dolev, J.H. Korhonen, C. Lenzen, J. Suomela and J. Rybicki. Synchronous counting and computational algorithm design. In *Proceedings of the 15th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2013)*, Osaka, Japan, November 13–16, 2013, Lecture Notes in Computer Science, vol. 8255, pp. 237-250.
11. M. Du, J. Kangasharju, O. Karkulahti, L. Pivovarova and R. Yangarber. Combined analysis of news and Twitter messages. In *Proceedings of the Joint Workshop on NLP&LOD and SWAIE: SemanticWeb, Linked Open Data and Information Extraction*, 2013, pp. 41-48.
12. A. Efrat, M. Nikkilä and V. Polishchuk. Sweeping a Terrain by Collaborative Aerial Vehicles. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL'13)*, 2013, pp. 4-13.
13. D. Entner, P.O. Hoyer and P. Spirtes. Data-driven covariate selection for non-parametric estimation of causal effects. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013, pp. 256-264.
14. F. Fomin, J.H. Korhonen and P. Golovach. On the Parameterized Complexity of Cutting a Few Vertices from a Graph. In *Proceedings of the 38th International Symposium on Mathematical Foundations of Computer Science 2013 (MFCS 2013)*, Klosterneuburg, Austria, August 26-30, 2013, Lecture Notes in Computer Science, vol. 8087, pp. 421-432.
15. T. Gagie. On the Value of Multiple Read/Write Streams for Data Compression. In H. Aydinian, F. Cicalese and C. Deppe (editors), *Information Theory, Combinatorics, and Search Theory: In Memory of Rudolf Ahlswede*, 2013, Lecture Notes in Computer Science, vol. 7777, pp. 284-297.
16. T. Gagie, P. Gawrychowski and Y. Nekrich. Heaviest Induced Ancestors and Longest Common Substrings. In *Proceedings of CCCG '13*, 2013.
17. T. Gagie, D. Hermelin, G.M. Landau and O. Weimann. Binary Jumbled Pattern Matching on Trees and Tree-Like Structures. In *Proceedings of the 21st Annual European Symposium on Algorithms (ESA 2013)*, Sophia Antipolis, France, September 2-4, 2013, Lecture Notes in Computer Science, vol. 8125, pp. 517-528.
18. T. Gagie, W.-H. Hon and T.-H. Ku. New Algorithms for Position Heaps. In *Proceedings of the 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013)*, Bad Herrenalb, Germany, June 17-19, 2013, Lecture Notes in Computer Science, vol. 7922, pp. 95-106.
19. T. Gagie, K. Karhu, G. Navarro, S. Puglisi and J. Siren. Document Listing on Repetitive Collections. In *Proceedings of the 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013)*, Bad Herrenalb, Germany, June 17-19, 2013, Lecture Notes in Computer Science, vol. 7922, pp. 107-119.

20. E. Giaquinta, K. Fredriksson, S. Grabowski and E. Ukkonen. Motif Matching Using Gapped Patterns. In *Revised Selected Papers of the 24th International Workshop on Combinatorial Algorithms (IWOCA 2013)*, Rouen, France, July 10-12, 2013, Lecture Notes in Computer Science, vol. 8288, pp. 448-452.
21. O. Gross, A. Doucet and H. Toivonen. Named Entity Filtering Based on Concept Association Graphs. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*, 2013.
22. S. Huttunen, A. Vihavainen, M. Du and R. Yangarber. Predicting Relevance of Event Extraction for the End User. In T. Poibeau, H. Saggion, J. Piskorski and R. Yangarber (editors), *Multi-source, Multilingual Information Extraction and Summarization, Theory and Applications of Natural Language Processing*, 2013, Springer-Verlag, Berlin, Heidelberg, pp. 163-176.
23. A. Hyttinen, P. Hoyer, F. Ederhardt and M. Järvisalo. Discovering Cyclic Causal Models with Latent Variables: A General SAT-Based Procedure. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013, pp. 301-310.
24. D. Ienco, A. Bifet, I. Zliobaite and B. Pfahringer. Clustering Based Active Learning for Evolving Data Streams. In *Proceedings of the 16th International Conference on Discovery Science*, 2013, pp. 79-93.
25. V. Jain and E. Galbrun. Topical organization of user comments and application to content recommendation. In *Proceedings of the 22nd International World Wide Web Conference (WWW '13)*, Companion Volume, 2013, pp. 61-62.
26. M. Kopotev and L. Pivovarova. Не до X": алгоритм выявления устойчивых параметров в сочетаниях слов. In Proceedings of the International Scientific Conference on Corpus Linguistics, 2013.
27. M. Kopotev, L. Pivovarova, N. Kochetkova and R. Yangarber. Automatic detection of stable grammatical features in n-grams. In *Proceedings of the 9th NAACL Workshop on Multiword Expressions (MWE 2013)*, 2013, Atlanta, United States.
28. J.H. Korhonen and P. Parviainen. Exact Learning of Bounded Tree-width Bayesian Networks. In *JMLR: Workshop and Conference Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, 2013, pp. 370–378.
29. A. Kotsifakos, P. Papapetrou and V. Athitsos. IBSM: Interval-Based Sequence Matching. In SDM, 2013.
30. J. Kärkkäinen, D. Kempa and S. Puglisi. Crochemore's String Matching Algorithm: Simplification, Extensions, Applications. In *Proceedings of the Prague Stringology Conference 2013*, 2013, pp. 168-175.
31. J. Kärkkäinen, D. Kempa and S. Puglisi. Lightweight Lempel-Ziv Parsing. In *Proceedings of the 12th International Symposium on Experimental Algorithms (SEA 2013)*, Rome, Italy, June 5-7, 2013, Lecture Notes in Computer Science, vol. 7933, pp. 139-150.
32. J. Kärkkäinen, D. Kempa and S. Puglisi. Linear Time Lempel-Ziv Factorization: Simple, Fast, Small. In *Proceedings of the 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013*), Bad Herrenalb, Germany, June 17-19, 2013, Lecture Notes in Computer Science, vol. 7922, pp. 189-200.
33. J. Kärkkäinen and G. Tischler. Near in Place Linear Time Minimum Redundancy Coding. In *Proceedings of the Data Compression Conference (DCC), 2013*, pp. 411-420.
34. A. Laaksonen. Efficient and simple algorithms for time-scaled and time-warped music search. In *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research*, 2013, pp. 621-630.
35. A. Laaksonen. Semi-automatic melody extraction using note onset time and pitch information from users. In *Proceedings of the Sound and Music Computing Conference 2013*, 2013, pp. 689-694.
36. A. Laaksonen and K. Lemström. On finding symbolic themes directly from audio using dynamic programming. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, 4-8 November, 2013, pp. 47-52.
37. J. Lijffijt. A fast and simple method for mining subsequences with surprising event counts. In H. Blockeel, K. Kersting, S. Nijssen and F. Zelezný (editors), *Machine Learning and Knowledge Discovery in Databases*, 2013, Springer Berlin Heidelberg, pp. 385-400.

38. S.M. Linkola, L. Rasku and T. Ahonen. Expandable String Representation for Music Features Using Adaptive-Size Alphabets. In *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research: Sound, Music & Motion*, 2013, pp. 920-927.
39. K. Mäki-Reinikka, J. Torniainen, A. Alafuzoff, H. Kotkanen and J.M. Toivanen. *Using Galvanic Vestibular Stimulation to Sense Abstract Data*. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology (ACE 2013)*, Enschede, Netherlands, Netherlands, 12-15 November, 2013.
40. T.M. Niinimäki and M. Koivisto. Annealed Importance Sampling for Structure Learning in Bayesian Networks. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013, pp. 1579-1585.
41. T.M. Niinimäki and M. Koivisto. Treedy: A Heuristic for Counting and Sampling Subset. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI-13)*, 2013, pp. 469-477.
42. J. Nouri, L. Pivovarova and R. Yangarber. MDL-based Models for Transliteration Generation. In *Proceedings of International Conference on Statistical Language and Speech Processing* (SLSP 2013), Tarragona, Spain, 29-31 July, 2013.
43. L. Pivovarova, M. Du and R. Yangarber. Adapting the PULS Event Extraction Framework to Analyze Russian Text. In Proceedings of *the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (ACL 2013)*, 2013, pp. 100-109.
44. L. Pivovarova, S. Huttunen and R. Yangarber. Event representation across genre. In *Proceedings of NAACL Workshop on EVENTS: Definition, Detection, Co-reference, and Representation*, Atlanta, GA, United States, 2013.
45. S. Puglisi and D. Kempa. Lempel-Ziv factorization: Simple, fast, practical. In *Proceedings of the Meeting on Algorithm Engineering and Experiments (ALENEX)*, 2013, pp. 103-112.
46. H. Su and J. Rousu. Multilabel Classification through Random Graph Ensembles. In *Proceedings of the 5th Asian Conference on Machine Learning (ACML2013)*, 2013, pp. 404-418.
47. N. Tatti. Itemsets for Real-valued Datasets. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM 2013)*, 2013.
48. N. Tatti and A. Gionis. Discovering Nested Communities. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013)*, 2013, pp. 32-47.
49. J. Toivanen, M. Järvisalo and H. Toivonen. Harnessing Constraint Programming for Poetry Composition. In *Proceedings of the Fourth International Conference on Computational Creativity*, 2013, pp. 160-167.
50. J. Toivanen, H. Toivonen and A. Valitutti. Automatical Composition of Lyrical Songs. In *Proceedings of the Fourth International Conference on Computational Creativity*, 2013.
51. J. Tuimala and A. Kallio. R, Programming Language. In *Encyclopedia of Systems Biology*, 2013, pp. 1809-1811.
52. A. Valitutti, A. Doucet, J. Toivanen and H. Toivonen. "Let Everything Turn Well in Your Wife": Generation of Adult Humor Using Lexical Constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 2: Short Papers*, 2013, pp. 243.
53. J. Wettig, J. Nouri, K. Reshetnikov and R. Yangarber. Information-theoretic modeling of etymological sound change. In L. Borin and A. Saxena (editors), *Approaches to measuring linguistic differences, Trends in Linguistics*, no. 265, 2013, Mouton de Gruyter, pp. 507-532.
54. I. Zliobaite and J. Hollmén. Fault tolerant regression for sensor data. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2013, pp. 449-464.

Technical reports and other publications

1. D.M. Hartley, N.P. Nelson, R.R. Arthur, P. Barboza, N. Collier, N. Lightfoot, J.P. Linge, E. van der Goot, A. Mawudeku, L.C. Madoff, L. Vaillant, R. Walters, R. Yangarber, J. Mantero, C.D. Corley

and J.S. Brownstein. An overview of Internet biosurveillance. In *Clinical Microbiology and Infection*, vol. 19, no. 11, 2013, pp. 1006-1013.
2. A. Hyvärinen. Independent component analysis: recent advances. In *Philosophical transactions - Royal Society, Mathematical, Physical and engineering sciences*, vol. 371, no. 1984, 2013
3. M. Shubin. Decomposing the Bacterial Phenotypic Time-series into biologically-meaningful components. In *Proceedings of the 11th Bioinformatics Research and Education Workshop*, Berlin, Germany, 3-4 May, 2013.
4. H. Toivonen, O. Gross, J.M. Toivanen and A. Valitutti. On Creative Uses of Word Associations. In *Advances in Intelligent Systems and Computing: Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, 2013, pp. 17-24.

Artistic work

1. O. Gross (editor) and H. Toivonen. Organization of Computational Creativity Club. 2013.
2. O. Holmqvist, D. Murphy and J. Paalasmaa. Biomusic performance, Gallery Jade, The Night of the Arts, 2013.
3. O. Holmqvist, D. Murphy and J. Paalasmaa. Biomusic performance, the 8th World Conference of Science Journalists, 2013.
4. S. Lääne, O. Gross, J. Toivanen and H. Toivonen. ArNePo: Arts, News & Poetry. 2013.
5. K. Mäki-Reinikka, J. Toivanen, A. Alafuzoff, H. Kotkanen and J. Torniainen. Brains on Art: Brain Poetry. 2013.
6. K. Mäki-Reinikka, J. Toivanen, A. Alafuzoff, H. Kotkanen and J. Torniainen. Brains on Art: The Suit. 2013.
7. K. Mäki-Reinikka, A. Alafuzoff, H. Kotkanen, J. Toivanen and J. Torniainen. Brains on Art: It's not just in your head. 2013.
8. H. Paakkanen and J. Toivanen. Oodimobiili. 2013.
9. H. Toivonen and T. Sivula. Soul Music: Making music of your dreams. 2013.
10. H. Toivonen and T. Sivula. Musical Chair. 2013.

2012

Articles in refereed scientific journals

1. A. Björklund, T. Husfeldt, P. Kaski and M. Koivisto. The traveling salesman problem in bounded degree graphs. *ACM Transactions on Algorithms*, 8 (2), Article 18, 2012, pp. 1-13.
2. F. Claude, G. Navarro, H. Peltola, L. Salmela and J. Tarhio. String matching with alphabet sampling. *Journal of Discrete Algorithms*, 11, 2012, pp. 37-50.
3. E. Congdon, S. Service, J. Wessman, J.K. Seppanen, S. Schönauer, J. Miettunen, H. Turunen, M. Koiranen, M. Joukamaa, M.-R. Jarvelin, L. Palotie, J. Veijola, H. Mannila, T. Paunio and N.B. Freimer. Early Environment and Neurobehavioral Development Predict Adult Temperament Clusters. *PLoS One*, 7(7), 2012, Article e38065.
4. J.T. Eronen, M. Fortelius, F. Portmann, K. Puolamäki and C.M. Janis. Neogene Aridification of the Northern Hemisphere. *Geology*, 40(9), 2012, pp. 823-826.
5. L. Eronen and H. Toivonen. Biomine: Predicting links between biological entities using network models of heterogeneous database. *BMC Bioinformatics*, 13, 2012, Article 119.
6. E. Galbrun and P. Miettinen. From black and white to full color: extending redescription mining outside the Boolean world. *Statistical analysis and data mining*, 5 (4), 2012, pp. 284-303.
7. M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13, 2012, pp. 307–361.
8. H. Heikinheimo, J.T. Eronen, A. Sennikov, C.D. Preston, E. Oikarinen, P. Uotila, H. Mannila and M. Fortelius. Convergence in the distribution patterns of Europe's plants and mammals is due to environmental forcing. *Journal of Biogeography*, 39(9), 2012, pp. 1633-1644.

9. M. Heinonen, H. Shen, N. Zamboni and J. Rousu. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, vol. 28, no. 18, 2012, pp. 2333-2341.
10. A. Hyttinen, F. Eberhardt and P.O. Hoyer. Learning Linear Cyclic Causal Models with Latent Variables. *Journal of Machine Learning Research*, vol. 13, 2012, pp. 3387-3439.
11. M. Kashif, S. Pietila, K. Artola, R.A.C. Jones, A.K. Tugume, V. Makinen and J.P.T. Valkonen. Detection of Viruses in Sweetpotato from Honduras and Guatemala Augmented by Deep-Sequencing of Small-RNAs. *Plant Disease*, vol. 96, no. 10, 2012, pp. 1430-1437.
12. J. Kim, J. S. B. Mitchell, V. Polishchuk, S. Yang and J. Zou. Routing Multi-Class Traffic Flows in the Plane. *Computational Geometry*, 45(3), 2012, pp. 99–114.
13. T. Kivioja, A.V. Vähärautio, K. Karlsson, A.W.M., Bonke, M. Enge, S. Linnarsson and J. Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9 (1), 2012, pp. 72–74.
14. A. Kotsifakos, P. Papapetrou, J. Hollmén, D. Gunopulos, V. Athitsos and G. Kollios. Hum-a-song: A Subsequence Matching with Gaps-Range-Tolerances Query-By-Humming System. *Proceedings of the Very Large Databases Endowement (PVLDB)*, 4(12), 2012, pp. 1930-1933.
15. I. Kostitsyna and V. Polishchuk. *Simple* Wriggling is Hard Unless You Are a Fat Hippo. *Theory of Computing Systems*, vol. 50, no. 1, 2012, pp. 93-110.
16. J. Lijffijt and S.Th. Gries. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 17(1), 2012, pp. 147-149.
17. L. Liu, K. Puolamäki, J.T. Eronen, M.M. Ataabadi, E. Hernesniemi and M. Fortelius. Dental functional traits of mammals resolve productivity in terrestrial ecosystems past and present. *Proceedings of the Royal Society B*, 279(1739), 2012, pp. 2793-2799.
18. H. Meronen, S.V. Henriksson, P. Raisanen and A. Laaksonen. Climate effects of northern hemisphere volcanic eruptions in an Earth System Model. *Atmospheric Research*, vol. 114, 2012, pp. 107-118.
19. V. Mäkinen, L. Salmela and J. Ylinen. Normalized N50 Assembly Metric using Gap-Restricted Co-Linear Chaining. *BMC Bioinformatics*, vol. 13, 2012, Article 255.
20. T. Niini, I. Scheinin, L. Lahti, S. Savola, F. Mertens, J. Hollmén, T. Böhling, A. Kivioja, K.H. Nord and S. Knuutila. Homozygous Deletions of Cadherin Genes in Chondrosarcoma — an Array CGH Study. *Cancer Genetics*, 205(11), 2012, pp. 588-593.
21. J. Pajula, J.-P. Kauppi and J. Tohka. Inter-Subject Correlation in fMRI: Method Validation against Stimulus-Model Based Analysis. *PLoS One*, vol. 7, no. 8, 2012, Article e41196.
22. P. Papapetrou, G. Benson and G. Kollios. Mining Poly-regions in DNA. *International Journal of Data Mining and Bioinformatics (IJDMB)*, 2012, INDERSCIENCE.
23. M.Á. Prada, J. Toivola, J. Kullaa and J. Hollmén. Three-way analysis of structural health monitoring data. *Neurocomputing*, 80, 2012, pp. 119–128.
24. P. Ramkumar, L. Parkkonen, R. Hari and A. Hyvärinen. Characterization of neuromagnetic brain rhythms over time scales of minutes using spatial independent component analysis. *Human Brain Mapping*, vol. 33, no. 7, 2012, pp. 1648-1662.
25. L. Salmela. Average complexity of backward q-gram string matching algorithms. *Information Processing Letters*, 112(11), 2012, pp. 433–437.
26. G. Sambuceti, M. Brignone, C. Marini, M. Massollo, F. Fiz, S. Morbelli, A. Buschiazzo, C. Campi, R. Piva, A.M. Massone, M. Piana and F. Frassoni. Estimating the whole bone-marrow asset in humans by a computational approach to integrated PET/CT imaging. *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 39, no. 8, 2012, pp. 1326-1338.
27. O. Solin, E. Ukkonen and L. Haikala. Mining the UKIDSS GPS: star formation and embedded clusters. *Astronomy & Astrophysics*, vol. 542, no 1, 2012, Article A3.
28. N. Välimäki, S. Ladra and V. Mäkinen. Approximate all-pairs suffix/prefix overlaps. *Information and Computation*, vol. 213, 2012, pp. 49-58.
29. J. Wessman, S. Schönauer, J. Miettunen, H. Turunen, P. Parviainen, J.K. Seppänen, E. Congdon, S. Service, M. Koiranen, J. Ekelund, J. Laitinen, A. Taanila, T. Tammelin, M. Hintsanen, L. Pulkki-

Råback, L. Keltikangas-Järvinen, J. Viikari, O.T. Raitakari, M. Joukamaa, M.-R. Järvelin, N. Freimer, L. Peltonen, J. Veijola, H. Mannila and T. Paunio. Temperament clusters in a normal population: implications for health and disease. *PLoS ONE*, vol. 7, no. 7, 2012, Article e33088.

Refereed conference articles and articles in edited books

1. P.R. Adhikari and J. Hollmén. Fast Progressive Training of Mixture Models for Model Selection. In J.-G. Ganascia, P. Lenca and J.-M. Petit (editors), *Proceedings of Fifteenth International Conference on Discovery Science (DS 2012)*, 2012, Springer-Verlag, pp. 194-208.
2. T. Ahonen. Compression-based Similarity Measuring in Music Information Retrieval. In *Proceedings of the Federated Computer Science Event 2012,* Helsinki, Finland, May 28-29, 2012, University of Helsinki, Department of Computer Science, Series of Publications B, Report B-2012-1, Helsinki, 2012, pp. 68-69.
3. T. E. Ahonen. Compression-based Clustering of Chromagram Data: New Method and Representations. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR 2012)*, London, UK, June, 2012, pp. 474-481.
4. E. Arkin, A. Efrat, G. Hart, I. Kostitsyna, A. Kröller, J. S. B. Mitchell and V. Polishchuk. Scandinavian Thins on Top of Cake: on the Smallest One-Size-Fits-All Box. In *Proceedings of the Sixth International Conference on Fun with Algorithms (FUN'12)*, Venice, Italy, June, 2012, pp. 16-27.
5. A. Björklund, T. Husfeldt, P. Kaski, M. Koivisto, J. Nederlof and P. Parviainen. Fast zeta transforms for point lattices. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2012)*, 2012, SIAM, pp. 1436-1444.
6. D. Entner and P. Hoyer. Estimating a Causal Order among Groups of Variables in Linear Models. In *Proceedings of the 22nd International Conference on Artificial Neural Networks, Artificial Neural Networks and Machine Learning (ICANN 2012)*, Lausanne, Switzerland, September 11-14, 2012, Lecture Notes in Computer Science, vol. 7553, 2012, pp. 83-90.
7. D. Entner, P.O. Hoyer and P. Spirtes. Statistical test for consistent estimation of causal effects in linear non-Gaussian models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS-2012)*, La Palma, Canary Islands, 2012, pp. 364-372.
8. J. Fischer, T. Gagie, T. Kopelowitz, M. Lewenstein, V. Mäkinen and N. Välimäki. Forbidden Patterns. In D. Fernández-Baca (editor), *Proceedings of 10th Latin American Symposium on Theoretical Informatics (LATIN 2012)*, Arequipa, Peru, April 16-20, 2012, Lecture Notes in Computer Science, Vol. 7256, Springer-Verlag, pp. 327–337.
9. T. Gagie, P. Gawrychowski, J. Kärkkäinen, Y. Nekrich and S.J. Puglisi. Faster Grammar-based Self-index. In *Proceedings of 6th International Conference on Language and Automata Theory and Applications (LATA 2012)*, A Coruña, Spain, March 5-9, 2012, pp. 240-251.
10. T. Gagie, K. Karhu, J. Kärkkäinen, V. Mäkinen, L. Salmela and J. Tarhio. Indexed Multi-Pattern Matching. In D. Fernández-Baca (editor), *Proceedings of 10th Latin American Symposium on Theoretical Informatics (LATIN 2012)*, Arequipa, Peru, April 16-20, 2012, Lecture Notes in Computer Science, Vol. 7256, Springer-Verlag, pp. 399-407.
11. E. Galbrun and A. Kimmig. Towards Finding Relational Redescriptions. In *Proceedings of the 15th International Conference on Discovery Science (DS 2012)*, Lyon, France, October 29-31, 2012, Lecture Notes in Computer Science, vol. 7569, 2012, pp. 52-66.
12. E. Galbrun and P. Miettinen. A Case of Visual and Interactive Data Analysis: Geospatial Redescription Mining. In *Proceedings of the ECML PKDD 2012 Workshop on Instant Interactive Data Mining*, Bristol, United Kingdom, September 24, 2012.
13. E. Galbrun and P.A. Miettinen. Siren: An Interactive Tool for Mining and Visualizing Geospatial Redescriptions [Demo]. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*, 2012, pp. 1544-1547.
14. S. Gaspers, M. Koivisto, M. Liedloff, S. Ordyniak and S. Szeider. On finding optimal polytrees. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI 2012)*, 2012, pp. 750-756.

15. C.D. Giurcaneanu, P. Luosto and P. Kontkanen. On the performance of histogram-based entropy estimators. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2012)*, 2012, pp. 1-6.
16. S. Gog, K. Karhu, J. Kärkkäinen, V. Mäkinen and N. Välimäki. Multi-Pattern Matching with Bidirectional Indexes. In *Proceedings of the 18th International Computing and Combinatorics Conference (COCOON 2012)*, Sydney, Australia, August 20-22, 2012, Lecture Notes in Computer Science, vol. 7434, 2012, pp. 384-395.
17. O. Gross, H. Toivonen, J. Toivanen and A. Valitutti. Lexical Creativity from Word Associations. In *Proceedings of the 7th IEEE International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2012)*, 2012, pp. 35-42.
18. M.U. Gutmann and A. Hyvärinen. Learning a selectivity-invariance-selectivity feature extraction architecture for images. *In Proceedings of International Conference on Pattern Recognition (ICPR 2012)*, Tsukuba, Japan, 2012, pp. 918-921.
19. M. Heinonen, N. Välimäki, V. Mäkinen and J. Rousu. Efficient Path Kernels for Reaction Function Prediction. In *Proceedings of the 3rd International Conference on Bioinformatics Models, Methods and Algorithms (Bioinformatics 2012)*, Algarve, Portugal, Feb 1-4, 2012, pp. 202-207.
20. J. Hirayama, A. Hyvärinen and S. Ishii. Structural equations and divisive normalization for energy-dependent component analysis. In *Advances in Neural Information Processing 25* (NIPS2011), vol. 24, Granada, Spain, 2012, pp. 1872-1880.
21. J. Hollmén. Mixture modeling of gait patterns in sensor data. In *Proceedings of the 5th International Conference on Pervasive Technologies and Relative to Assistive Environments (PETRA 2012)*, Crete, Greece, June 6-8, 2012, ACM, Article no 48.
22. A. Hyttinen, F. Eberhardt and P. Hoyer. Causal Discovery of Linear Cyclic Models from Multiple Experimental Data Sets with Overlapping Variables. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI'12)*, 2012, pp. 387-396.
23. M. Järvisalo, P. Kaski, M. Koivisto and J.H. Korhonen. Finding Efficient Circuits for Ensemble Computation. In *Proceedings of the 15th International Conference on Theory and Applications of Satisfiability Testing (SAT 2012)*, Trento, Italy, June 17-20, 2012, Lecture Notes in Computer Science, vol. 7317, 2012, pp. 369-382.
24. P. Kaski, M. Koivisto and J.H. Korhonen. Fast Monotone Summation over Disjoint Sets. In *Proceedings of the 7th International Symposium on Parameterized and Exact Computation (IPEC 2012)*, Ljubljana, Slovenia, September 12-14, 2012, Lecture Notes in Computer Science, vol. 7535, 2012, pp. 159-170.
25. A. Kotsifakos, P. Papapetrou, J. Hollmén, D. Gunopulos and V. Athitsos. A survey of query-by-humming similarity methods. In *Proceedings of the 5th International Conference on Pervasive Technologies and Relative to Assistive Environments (PETRA 2012)*, Crete, Greece, June 6-8, 2012, ACM, Article no 5.
26. J. Kärkkäinen, D. Kempa and S.J. Puglisi. Slashing the Time for BWT Inversion. In *Proceedings of 2012 Data Compression Conference (DCC 2012)*, IEEE Computer Society, pp. 99-108.
27. J. Kärkkäinen, P. Mikkola and D. Kempa. Grammar Precompression Speeds Up Burrows-Wheeler Compression. In *Proceedings of the 19th International Symposium on String Processing and Information Retrieval (SPIRE 2012)*, Cartagena de Indias, Colombia, October 21-25, 2012, Lecture Notes in Computer Science, vol. 7608, 2012, pp. 330-335.
28. L.A. Langohr and H. Toivonen. A Model for Mining Relevant and Non-redundant Information. In *Proceedings of the 27th ACM Symposium on Applied Computing (SAC 2012)*, Trento, Italy, 2012, pp. 132-137.
29. L.A. Langohr and H. Toivonen. Retrieval of Relevant and Non-redundant Nodes. In *Proceedings of The First SDM Workshop on Dynamic Network Analysis, in conjunction with Twelfth SIAM International Conference on Data Mining*, Anaheim, California, USA, April 2012, pp. 32-40.
30. J. Lijffijt, P. Papapetrou and K. Puolamäki. Size Matters: Finding the Most Informative Set of Window Lengths. In P.A. Flach, T. De Bie and N. Cristianini (editors), *Machine Learning and Knowledge Discovery in Databases*, 2012, Springer Berlin Heidelberg, pp. 451-466.

31. J. Lijffijt, T. Säily and T. Nevalainen. CEECing the baseline: Lexical stability and significant change in a historical corpus. In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen and Matti Rissanen (editors), *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources.* Research Unit for Variation, Contacts and Change in English, Helsinki, 2012.
32. P. Luosto, C.D. Giurcaneanu and P. Kontkanen. Construction of irregular histograms by penalized maximum likelihood: a comparative study. In *Proceedings of the IEEE Information Theory Workshop (ITW'12)*, 2012, pp. 297-301.
33. T. Niinimäki and P. Parviainen. Local Structure Discovery in Bayesian Networks. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI-12)*, 2012, pp. 634-643.
34. J. Paalasmaa, D.J. Murphy and O.Holmqvist. Analysis of Noisy Biosignals for Musical Performance. In *Proceedings of the 11th International Symposium on Advances in Intelligent Data Analysis XI (IDA 2012)*, Helsinki, Finland, October 25-27, 2012, Lecture Notes in Computer Science, vol. 7619, 2012, pp. 241–252.
35. J. Paalasmaa, M. Waris, H. Toivonen, L. Leppäkorpi and M. Partinen. Unobtrusive Online Monitoring of Sleep at Home. *In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'12)*, 2012, pp. 3784-3788.
36. P. Papapetrou, T. Chistiakova, J. Hollmén, V. Kalogeraki and D. Gunopulos. Finding representative objects using link analysis ranking. In *Proceedings of the 5th International Conference on Pervasive Technologies and Relative to Assistive Environments (PETRA 2012)*, Crete, Greece, June 6-8, 2012, ACM, Article no 6.
37. V. Polishchuk and T. Andersson. Socially optimal allocation of ATM resources via truthful market-based mechanisms. In *Proceedings of the 2nd SESAR Innovation Days (SID 2012)* , 2012.
38. S. Sankararaman, K. Abu-affash, A. Efrat, S.D. Eriksson-Bique, V. Polishchuk, S. Ramasubramanian and M. Segal. Optimization Schemes for Protective Jamming. In *Proceedings of the Thirteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2012)*, Hilton Head Island, NC, United States, 2012, pp. 65-74.
39. H. Sasaki, M.U. Gutmann, H. Shouno and A. Hyvärinen. *Topographic Analysis of Correlated Components.* In *Proceedings of the Asian Conference on Machine Learning, Journal of Machine Learning Research – Proceedings Track*, vol. 25, 2012, pp. 365-378.
40. M.E. Skarkala, M. Maragoudakis, S. Gritzalis, L. Mitrou, H. Toivonen and P. Moen. Privacy Preservation by k-Anonymization of Weighted Social Networks. Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), Istanbul, Turkey, August 26-29, 2012, pp. 423-428.
41. H. Su and J. Rousu. Random Graph Ensembles in Multilabel Classification. In *ICML 2012 Workshop, Object, Functional and Structured data: towards Next Generation Kernel-based Methods*, 2012.
42. T. Tashiro, S. Shimizu, A. Hyvärinen and T. Washio. Estimation of causal orders in a linear non-gaussian acyclic model: a method robust against latent confounders. In *Proceedings of 22nd International Conference on Artificial Neural Networks: the Artificial Neural Networks and Machine Learning (ICANN 2012)*, Lausanne, Switzerland, September 11-14, 2012, Lecture Notes in Computer Science, vol. 7552, 2012, pp. 491-498.
43. M.J. Timonen. Categorization of Very Short Documents. In *Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2012)*, 2012, pp. 5-16.
44. J. Toivanen, H. Toivonen, A. Valitutti and O. Gross. Corpus-based generation of content and form in poetry. In *Proceedings of International Conference on Computational Creativity (ICCC)*, Dublin, Ireland, May-June 2012, pp. 175-179.
45. A. Tulilaulu, J. Paalasmaa, M. Waris and H. Toivonen. Sleep Musicalization: Automatic Music Composition from Sleep Measurements. In *Proceedings of the 11th International Symposium on Advances in Intelligent Data Analysis XI (IDA 2012)*, Helsinki, Finland, October 25-27, 2012, Lecture Notes in Computer Science, vol. 7619, 2012, pp. 392-403.
46. T. Vartiainen and J. Lijffijt. Premodifying -ing participles in the parsed BNC. In J. Mukherjee and M. Huber (editors), *Corpus Linguistics and Variation in English: Theory and Description*, Rodopi Amsterdam, 2012, pp. 247-258.

47. A. Valitutti. Ambiguous Lexical Resources for Computational Humor Generation. In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence (ICAART)*, Vilamoura, Algrave, Portugal, February 6-8, 2012, SciTePress, pp. 532-535.
48. N. Välimäki. Least Random Suffix/Prefix Matches in Output-Sensitive Time. In *Proceedings of the 23rd Annual Symposium on Combinatorial Pattern Matching (CPM 2012)*, Helsinki, Finland, July 3-5, 2012, Lecture Notes in Computer Science, vol. 7354, 2012, pp. 269-279.
49. N. Välimäki and S.J. Puglisi. Distributed String Mining for High-Throughput Sequencing Data. In *Proceedings of the 12th International Workshop on Algorithms in Bioinformatics (WABI 2012)*, Ljubljana, Slovenia, September 10-12, 2012, Lecture Notes in Computer Science, vol. 7534, 2012, pp. 441-452.
50. J. Wettig, J. Nouri, K. Reshetnikov and R. Yangarber. Information-theoretic Methods for Analysis and Inference in Etymology. In *Proceedings of the Fifth Workshop on Information-theoretic Methods in Science and Engineering*, 2012, pp. 53-56.
51. H. Wettig, K. Reshetnikov and R. Yangarber. Using Context and Phonetic Features in Models of Etymological Sound Change. In *Proceedings of EACL 2012 Joint: Workshop on Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, Avignon, France, 2012, pp. 108-116.

Technical reports and other publications

1. L. Eronen, P. Hintsanen and H. Toivonen. Biomine: A network-structured resource of biological entities for link prediction. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), vol. 7250, Springer-Verlag, 2012, pp. 364–378.
2. O. Gross, A. Doucet and H. Toivonen. Term Association Analysis for Named Entity Filtering. In *Proceedings of The Twenty First Text REtrieval Conference (TREC 2012)*, Gaithersburg, Maryland, November 6-9, 2012, pp. 67.
3. S.E. Heinonen, E. Koivisto, M. Heinonen and P. Heikkinen. Seismic 3D modelling: case studies from Pyhäsalmi and Kevitsa, Finland. In *Proceedings of the Seventh Symposium on the Structure, Composition and Evolution of the Lithosphere in Finland (LITHOSPHERE 2012),* Institute of Seismology Report, no. S-56, pp. 25-28.
4. S.M. Huttunen. Tietojenkäsittelytiede: Tiedoneristäminen. V. Heikkinen, E. Voutilainen, P. Lauerma, U. Tiililä and M. Lounela (editors), *Genreanalyysi – tekstilajitutkimuksen käsikirja*. Kotimaisten kielten keskuksen julkaisuja, no. 169, Gaudeamus Helsinki University Press, Helsinki, 2012, pp. 696-706.
5. T. Hynönen, S.J. Mahler and H. Toivonen. Discovery of Novel Term Associations in a Document Collection. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, pp. 91–103.
6. A. Kimmig, E. Galbrun, H. Toivonen and L. De Raedt. Patterns and Logic for Reasoning with Networks. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, pp. 122-143.
7. J. Kärkkäinen and J. Stoye, editors. *Proceedings of the 23rd Annual Symposium on Combinatorial Pattern Matching (CPM 2012*), Helsinki, Finland, July 3-5, 2012, Lecture Notes in Computer Science, no. 7354, Springer-Verlag, 2012.
8. L.A. Langohr, V. Podpecan, M. Petek, I. Mozetic and K. Gruden. Contrast Mining from Interesting Subgroups. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, pp. 390-406.
9. L.A. Langohr and H. Toivonen. Finding representative nodes in probabilistic graphs. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools,*

and Applications, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012,, pp. 218–229.
10. P. Luosto, P. Kontkanen and K. Luosto. *The normalized maximum likelihood distribution of the multinomial model class with positive maximum likelihood parameters*. Department of Computer Science, Series of Publications C, no. C-2012-5, University of Helsinki, Department of Computer Science, Helsinki, 2012.
11. A. Mozetic, N. Lavrac, V. Podpecan, P.K. Novak, H. Motaln, M. Petek, K. Gruden, H. Toivonen and K. Kulovesi. Semantic Subgroup Discovery and Cross-context Linking for Microarray Data Analysis. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, pp. 379-389.
12. J. Piskorski and R. Yangarber. Information Extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization: Theory and Applications of Natural Language Processing*, Springer-Verlag, 2012, pp. 23-49.
13. T. Poibeau, H. Saggion, J. Piskorski, R. Yangarber, editors. *Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing, Springer-Verlag, 2012*, Berlin, Heidelberg, 2012.
14. J. Toivanen and P. Toivanen. Konedadaa vai keskeisloruja? *Parnasso*, no. 4, 2012, pp. 22–25.
15. H. Toivonen. Network Analysis: Overview. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, pp. 144–146.
16. H. Toivonen, F. Zhou, A. Hartikainen and A.E. Hinkka. Network Compression by Node and Edge Mergers. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, pp. 199–217.
17. A. Valitutti. Creative Coding for Humor Design: A Preliminary Exploration. *Extended abstracts of the 3rd International Workshop on Computational Humor (Computational Humor 2012)*, 2012, pp. 39-40.
18. A. Valitutti, H. Toivonen, O. Gross and J. Toivanen. Decomposition and Distribution of Humorous Effect in Interactive Systems. In *Proceedings of the AAAI Fall Symposium Series on Artificial Intelligence of Humor*, AAAI Technical Reports, no. FS-12-02, 2012, pp. 96-100.
19. F. Zhou, S.J. Mahler and H. Toivonen. Review of Network Abstraction Techniques. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, pp. 166–178.
20. F. Zhou, S.J. Mahler and H. Toivonen. Simplification of Networks by Edge Pruning. In M.R. Berthold (editor), *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, pp. 179–198.

Artistic work

1. H. Paakkanen and J. Toivanen. *Tee se kotona*. Kiitos. 2012.

2011

Articles in refereed scientific journals

1. S. Angelov, S. Inenaga, T. Kivioja and V. Mäkinen. Missing pattern discovery. *Journal of Discrete Algorithms*, 9(2), 2011, pp. 153-165.
2. A. Apostolico, C. Pizzi and E. Ukkonen. Efficient algorithms for the discovery of gapped factors. *Algorithms for Molecular Biology*, 6(5), 2011.

3. E. Arkin, M. Bender, J. Mitchell and V. Polishchuk. The Snowblower Problem. *Computational Geometry*, 44(8), 2011, pp. 370-384.
4. K. Astikainen, L. Holm, E. Pitkänen, S. Szedmak and J. Rousu. Structured Output Prediction of Novel Enzyme Function with Reaction Kernels. In *Biomedical Engineering Systems and Technologies Communications of Computer and Information Science*, 127 (5), 2011, pp. 367-378.
5. A. Björklund, T. Husfeldt, P. Kaski and M. Koivisto. Covering and packing in linear space. *Information Processing Letters*, 111(21-22), 2011, pp. 1033-1036.
6. J.T. Eronen, K. Puolamäki, H. Heikinheimo, H. Lokki, A. Venäläinen, H. Mannila and M. Fortelius. The effect of scale, climate and environment on species richness and spatial distribution of Finnish birds. *Annales Zoologici Fennici*, 48(5), 2011, pp. 257-274.
7. P. Floréen, M. Hassinen, J. Kaasinen, P. Kaski, T. Musto and J. Suomela. Local approximability of max-min and min-max linear programs. *Theory of Computing Systems*, 49(4), 2011, pp. 672-697.
8. G.C. Garriga, E. Junttila and H. Mannila. Banded structure in binary matrices. *Knowledge and Information Systems*, 28(1), 2011, pp. 197-226.
9. R. Grote, J.H. Korhonen and I. Mammarella. Challenges for process-based modelling of gas exchange in mixed forests. *Forest Systems*, 20(3), 2011, pp. 389-406.
10. M. Hassinen, J. Kaasinen, E. Kranakis, V. Polishchuk, J. Suomela and A. Wiese. Analysing local algorithms in location-aware quasi-unit-disk graphs. *Discrete Applied Mathematics*, 159(15), 2011, pp. 1566-1580.
11. M. Heinonen, S. Lappalainen, T.J. Mielikäinen and J. Rousu. Computing Atom Mappings for Biochemical Reactions without Subgraph Isomorphism. *Journal of Computational Biology*, 18(1), 2011, pp. 43-58.
12. A. Hulpke, P. Kaski and P.R.J. Östergård. The number of Latin squares of order 11. *Mathematics of Computation*, 80, 2011, pp. 1197-1219.
13. A. Hyvärinen. Testing the ICA mixing matrix based on inter-subject or inter-session consistency. *NeuroImage*, 58(1), 2011, pp. 122-136.
14. A. Kallio, K. Puolamäki, M. Fortelius and H. Mannila. Correlations and co-occurrences of taxa: the role of temporal, geographic, and taxonomic restrictions. *Palaeontologia Electronica*, 14(1), 2011, pp. 4A.
15. A. Kallio, N. Vuokko, M. Ojala, N. Haiminen and H. Mannila. Randomization techniques for assessing the significance of gene periodicity results. *BMC Bioinformatics*, 12, 2011, pp. 330.
16. P. Kaski, V. Mäkinen and P.R.J. Östergård. The Cycle Switching Graph of the Steiner Triple Systems of Order 19 is Connected. *Graphs and Combinatorics*, 27(4), 2011, pp. 539-546.
17. M. Korpela, P. Nöjd, J. Hollmén, H. Mäkinen, M. Sulkava and P. Hari. Photosynthesis, temperature and radial growth of Scots pine in northern Finland: identifying the influential time intervals. *Trees - Structure and Function*, 25(2), April 2011, pp. 323-332.
18. A. Kotsifakos, P. Papapetrou, J. Hollmén and D. Gunopulos. A subsequence matching with gaps-range-tolerances framework: A query-by-humming application. *Proceedings of the VLDB Endowment*, 4(11), 2011, pp. 761-771.
19. P. Luosto and P. Kontkanen. Clustgrams: an extension to histogram densities based on the minimum description length principle. *Central European Journal of Computer Science*, 1(4), 2011, pp. 466-481.
20. T. Nevalainen, H. Raumolin-Brunberg and H. Mannila.The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change. *Language Variation and Change*, 23, 2011, pp. 1-43.
21. P. Papapetrou, V. Athitsos, M. Potamias, G. Kollios and D. Gunopulos. Embedding-based Subsequence Matching in Time Series Databases. In *ACM Transactions on Database Systems (TODS)*, 36(3), 2011, pp. 17.
22. A. Pizzi, P. Rastas and E. Ukkonen. Finding significant matches of position weight matrices in linear time. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1), 2011, pp. 69-79.

23. V. Podpecan, N. Lavrac, I. Mozetic, P. Kralj Novak, I. Trajkovski, L. Langohr, K. Kulovesi, H. Toivonen, M. Petek, H. Motaln and K. Gruden. SegMine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinformatics*, 12(416), 2011.
24. A. Rizo, K. Lemström and J.M. Iñesta. Polyphonic music retrieval with classifier ensembles. *Journal of New Music Research*, 40(4), 2011, pp. 313-325.
25. L. Salmela, V. Mäkinen, N. Välimäki, J. Ylinen and E. Ukkonen. Fast scaffolding with small independent mixed integer programs. *Bioinformatics*, 27(23), 2011, pp. 3259-3265.
26. L. Salmela and J. Schröder. Correcting errors in short reads by multiple alignments. *Bioinformatics*, 27(11), 2011, pp. 1455-1461.
27. S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. Hoyer and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research* 12, 2011.
28. Y. Sogawa, S. Shimizu, A. Hyvärinen, T. Washio, T. Shimamura and S. Imoto. Estimating Exogenous Variables in Data with More Variables than Observations. *Neural Networks*, 24(8), 2011, pp. 875-880.
29. M. Sulkava, S. Luyssaert, S. Zaehle and D. Papale. Assessing and improving the representativeness of monitoring networks: The European flux tower network example. *Journal of Geophysical Research - Biogeosciences*, 116, May 2011, pp. G00J04.
30. P. Virtala, V. Berg, M.K. Kivioja, J. Purhonen, M. Salmenkivi, P. and M. Tervaniemi. *T*he preattentive processing of major vs. minor chords in the human brain. An event-related potential study. *Neuroscience Letters*, 487(3), 2011, pp. 406-410.
31. S. Yang, J. Mitchell, J. Krozel, V. Polishchuk, J. Kim and J. Zou. Flexible Airlane Generation to Maximize Flow under Hard and Soft Constraints. *Air Traffic Control Quarterly*, 19(3), 2011, pp. 1-26.

Refereed conference articles and articles in edited books

1. P. R. Adhikari, B.B. Upadhyaya, C. Meng and J. Hollmén. Gene selection in time-series gene expression data. In M. Loog, L. Wessels, M.J.T. Reinders, and D. de Ridder (editors), *Proceedings of the 6th IAPR Conference on Pattern Recognition in Bioinformatics*, November, 2011, Lecture Notes in Bioinformatics, Vol. 7036, Springer-Verlag, pp. 145-156.
2. P. Agarwal, A. Efrat, C. Gniady, J. Mitchell, V. Polishchuk and G. Sabhnani. Distributed Localization and Clustering Using Data Correlation and the Occam's Razor Principle. In *Proceedings of 2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS)*, 2011.
3. T. Ahonen. Kolmogorov Complexity in Lyrics. In *Proceedings of AdMIRe 2011*.
4. T. Ahonen, K. Lemström and S.M. Linkola. Compression-based Similarity Measures in Symbolic, Polyphonic Music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*,Miami, Usa, October, 2011, pp. 91-96.
5. S. Alonso, M. Dominguez, M. A. Prada, M. Sulkava and J. Hollmén. Comparative analysis of power consumption in university buildings using envSOM. In J. Gama, E. Bradley and J. Hollmén (editors), *Proceedings of 10th International Symposium on Advances in Intelligent Data Analysis (IDA 2011)*, Porto, Portugal, October, 2011, Lecture Notes in Computer Science, Vol. 7014, Springer-Verlag, pp. 10-21.
6. S. Alonso, M. Sulkava, M.A. Prada, M. Dominguez, and Jaakko Hollmén. EnvSOM: a SOM algorithm conditioned on the environment for clustering and visualization. In J. Laaksonen and T. Honkela (editors), *Proceedings of 8th International Conference on Advances in Self-Organizing Maps (WSOM 2011)*, Espoo, Finland, June, 2011, Aalto University, Lecture Notes in Computer Science, Vol. 6731, Springer-Verlag, pp. 61-70.
7. E. Arkin, C. Dieckmann, C. Knauer, J. Mitchell, V. Polishchuk, L. Schlipf and S. Yang. Convex Transversals. In *Proceedings of 12th International Symposium on Algorithms and Data Structures (WADS 2011)*, New York, NY, USA, August 15-17, 2011, Lecture Notes in Computer Science, Vol. 6844, Springer-Verlag, pp. 49-60.

8. M. Atkinson, J. Piskorski, E. Van der Goot and R. Yangarber. Multilingual real-time event extraction for border security intelligence gathering. In Uffe Kock Wiil (editor), *Counterterrorism and Open Source Intelligence,* Lecture Notes in Social Networks, Vol. 2, Springer-Verlag, 2011, pp. 355-390.
9. M.J. Brewer, M. Sulkava, H. Mäkinen, M. Korpela, P. Nöjd and J. Hollmén. Logistic fitting method for detecting onset and cessation of tree stem radius increase. In H. Yin, W. Wang and V. Rayward-Smith (editors), *Proceedings of 12th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2011)*, Norwich, UK, September 2011, Lecture Notes in Computer Science, Vol. 6936, Springer-Verlag, pp. 204-211.
10. M. Du, P. von Etter, M. Kopotev, M. Novikov, N. Tarbeeva and R. Yangarber. Building support tools for Russian-language information extraction. In *Proceedings of Balto-Slavonic Natural Language Processing (BSNLP-2011)*, Plzeň, Czech Republic, 2011.
11. D. Entner and P.O. Hoyer. Discovering Unconfounded Causal Relationships Using Linear Non-Gaussian Models. In *Proceedings of New Frontiers in Artificial Intelligence: JSAI-isAI 2010 Workshops*, Tokyo, Japan, November 18-19, 2010, Lecture Notes in Computer Science, Vol. 6797, Revised Selected Papers, Springer-Verlag, 2011, pp. 181-195.
12. P. Ferragina, J. Sirén and R. Venturini. Distribution-Aware Compressed Full-Text Indexes. In *Proceedings of the 19th Annual European Symposium on Algorithms (ESA 2011)*, Saarbrücken, Germany, September, 2011, Lecture Notes in Computer Science, Vol. 6942, Springer-Verlag, pp. 760-771.
13. E. Galbrun and P. Miettinen. From Black and White to Full Colour: Extending Redescription Mining Outside the Boolean World. In *Proceedings of SIAM International Conference on Data Mining*.
14. M.U. Gutmann and A. Hyvärinen. Extracting coactivated features from multiple datasets. In *Proceedings of 21st International Conference on Artificial Neural Networks (ICANN 2011)*, Espoo, Finland, June 14-17, 2011, Lecture Notes in Computer Science, Vol. 6791, Springer-Verlag, pp. 323-330.
15. M. Gutmann and J. Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, Barcelona, Spain, 2011, pp. 283-290.
16. S. Huttunen, A. Vihavainen, P. von Etter and R.Yangarber. Relevance prediction in information extraction using discourse and lexical features. In *Proceedings of Nordic Conference on Computational Linguistics (Nodalida-2011)*, Riga, Latvia, 2011.
17. A. Hyttinen, F. Eberhardt and P. Hoyer. Noisy-OR Models with Latent Confounding. In *Proceedings of the twenty-seventh conference on Uncertainty in Artificial Intelligence (UAI-11)*, 2011, pp. 363-372.
18. E. Junttila and P. Kaski. Segmented nestedness in binary data. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM'11)*, Mesa, Arizona, USA, 28-30 April, 2011, SIAM/Omnipress, pp. 235-246.
19. M. Karvonen, M. Laitinen, K. Lemström and J. Vikman. Error-Tolerant Content-Based Music-Retrieval with Mathematical Morphology. In *Proceedings of 7th International Symposium on Exploring Music Contents (CMMR 2010)*, Málaga, Spain, June 21-24, 2010, Lecture Notes in Computer Science, Vol. 6684, Revised Papers, Springer-Verlag, 2011, pp. 321-337.
20. M. Kopotev, M. Du, P. von Etter, M. Novikov, N. Tarbeeva and R. Yangarber. Building Support Tools for Russian-Language Information Extraction. In *Proceedings of 14th International Conference on Text, Speech and Dialogue (TSD 2011)*, Pilsen, Czech Republic, September 1-5, 2011, Lecture Notes in Computer Science, Vol. 6836, Springer-Verlag, pp. 380-387.
21. O. Kostakis, P. Papapetrou and J. Hollmén. ARTEMIS: Assessing the similarity of event-interval sequences. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis (editors), *Proceedings of the Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD 2011)*, September, 2011, Lecture Notes in Computer Science, Vol. 6912, Springer-Verlag, pp. 229-244.

22. O. Kostakis, P. Papapetrou and J. Hollmén. Distance measure for querying arrangements of temporal intervals. In *Proceedings of 4th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2011)*, Crete, Greece, May, 2011, ACM.
23. A. Kotsifakos, V. Athitsos, P. Papapetrou, J. Hollmén and D. Gunopulos. Model-based search in large time series databases. In *Proceedings of The 4th International Conference on Pervasive Technologies Related to Assistive Environment (PETRA 2011)*, Crete, Greece, May 2011, ACM.
24. J. Krozel, M. Ganji, S. Yang, J. Mitchell and V. Polishchuk. Metrics for evaluating the impact of weather on jet routes. In *Proceedings of 15th Conference on Aviation, Range, and Aerospace Meteorology*, 2011.
25. J. Krozel, S. Yang, J. Mitchell and V. Polishchuk. Strategies to Mitigate Off-Nominal Events in Super Dense Operations. In *Proceedings of AIAA Guidance, Navigation, and Control Conference*, 2011.
26. J. Kärkkäinen and T. Gagie. Counting Colours in Compressed Strings. In *Proceedings of 22nd Annual Symposium on Combinatorial Pattern Matching (CPM 2011)*, Palermo, Italy, June 27-29, 2011, Lecture Notes in Computer Science, vol. 6818, Springer-Verlag, pp. 197-207.
27. J. Kärkkäinen and S.J. Puglisi. Fixed Block Compression Boosting in FM Indexes. In *Proceedings of 18th International Symposium on String Processing and Information Retrieval (SPIRE 2011)*, Pisa, Italy, October 17-21, 2011, Lecture Notes in Computer Science, Vol. 7024, Springer-Verlag, pp. 174-184.
28. J. Kärkkäinen and S.J. Puglisi. Cache-Friendly Burrows-Wheeler Inversion. In *Proceedings of First International Conference on Data Compression, Communications and Processing (CCP 2011)*, 2011, pp. 38-42.
29. V. Laparra, M.U. Gutmann, J. Malo and A. Hyvärinen. Complex-valued independent component analysis of natural images. In *Proceedings of 21st International Conference on Artificial Neural Networks (ICANN 2011)*, Espoo, Finland, June 14-17, 2011, Lecture Notes in Computer Science, vol. 6792, Springer-Verlag, pp. 213-220.
30. M. Laitinen and K. Lemström. Dynamic Programming in Transposition and Time-Warp Invariant Polyphonic Content-Based Music Retrieval. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11)*, Miami, Florida, USA, October 2011, pp. 369-374.
31. K. Lemström and M. Laitinen. Transposition and time-warp invariant geometric music retrieval algorithms. In *Proceedings of 2011 IEEE International Conference on Multimedia and Expo (ICME)*, Barcelona, Spain, 2011, pp. 1-6.
32. J. Lijffijt, P. Papapetrou, K. Puolamäki and H. Mannila. Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In *Proceedings of the European conference of Machine learning and knowledge discovery in databases - Part II*, 2011, SpringerVerlag, pp. 341-357.
33. A. Moneta, N. Chlaß, D. Entner and P.O. Hoyer. Causal Search in Structural Vector Autoregressive Models. In *Proceedings of NIPS Mini-Symposium on Causality in Time Series*, 2011, pp. 95-118.
34. T.M. Niinimäki, P. Parviainen and M. Koivisto. Partial Order MCMC for Structure Discovery in Bayesian Networks. In *Proceedings of the Twenty-Seventh Conference Conference on Uncertainty in Artificial Intelligence (UAI-11)*, 2011, AUAI Press, pp. 557-564.
35. J. Paalasmaa, L. Leppäkorpi and M. Partinen. Quantifying respiratory variation with force sensor measurements. In *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'11)*, Boston, USA, 2011, pp. 3812-3814.
36. P. Papapetrou, A. Gionis and H. Mannila. A Shapley-value Approach for Influence Attribution. In *Proceedings of the European Conference of Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML PKDD)*, Athens, Greece, September 5-9, 2011, Lecture Notes in Computer Science, Vol. 6912, Springer-Verlag, pp. 549-564.
37. P. Parviainen and M. Koivisto. Ancestor Relations in the Presence of Unobserved Variables. In *Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2011)*, Athens, Greece, September 5-9, 2011, Lecture Notes in Computer Science, Vol. 6912, Springer-Verlag, pp. 581-596.

38. E. Pitkänen, M. Arvas and J. Rousu. Minimum mutation algorithm for gapless metabolic network evolution. In *Proceedings of International Conference of Bioinformatics Models, Methods and Algorithms (Bioinformatics 2011)*, Rome, Italy, January, 2011, pp. 28-38.
39. V. Polishchuk and M.J. Sysikaski. Faster algorithms for minimum-link paths with restricted orientations. In *Proceedings of 12th International Symposium on Algorithms and Data Structures (WADS 2011)*, New York, NY, USA, August 15-17, 2011, Lecture Notes in Computer Science, vol. 6844, Springer-Verlag, pp. 655-666.
40. J.S. Puuronen and A. Hyvärinen. Hermite Polynomials and Measures of Non-Gaussianity. In *Proceedings of 21st International Conference Artificial Neural Networks (ICANN2011)*, 2011, pp. 205-212.
41. E. Rivals, L. Salmela and J. Tarhio. Exact search algorithms for biological sequences. In M. Elloumi and A.Y. Zomaya (editors), *Algorithms in computational molecular biology: Techniques, approaches and applications, Bioinformatics: Computational Techniques and Engineering*, 2011, John Wiley & Sons, pp. 91-111.
42. J. Rousu, D. Agranoff, J. Shawe-Taylor and D. Fernandez-Reyes. Sparse Canonical Correlation Analysis for Biomarker Discovery: A Case Study in Tuberculosis. In *Proceedings of the Fifth International Workshop on Machine Learning in Systems Biology*, 2011, pp. 73-77.
43. J. Rousu and H. Su. Multi-Task Drug Bioactivity Classification with Graph Labeling Ensembles. In *Proceedings of the 6th International Conference on Pattern Recognition in Bioinformatics*, Delft, The Netherlands, November, 2011, Lecture Notes in Computer Science, Vol. 7036, Springer-Verlag, pp. 157-167.
44. H. Sasaki, M.U. Gutmann, H. Shouno and A. Hyvärinen. Learning Topographic Representations for Linearly Correlated Components. In *Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
45. J. Sirén, N. Välimäki and V. Mäkinen. Indexing Finite Language Representation of Population Genotypes. In *Proceedings of 11th International Workshop on Algorithms in Bioinformatics (WABI 2011)*, Saarbrücken, Germany, September, 2011, Lecture Notes in Bioinformatics, Vol. 6833, pp. 270-281.
46. J. Toivola and J. Hollmén. Collaborative filtering for coordinated monitoring in sensor networks. In *Proceedings of the ICDMW 2011 11th IEEE International Conference on Data Mining Workshops*, Vancouver, Canada, December, 2011, IEEE Computer Society, pp. 987-994.
47. H. Toivonen, F. Zhou, A. Hartikainen and A. Hinkka. Compression of Weighted Graphs. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, San Diego, USA, August, 2011, pp. 965-973.
48. A. Valitutti. How Many Jokes are Really Funny? Towards a New Approach to the Evaluation of Computational Humour Generators. In *Proceedings of International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2011)*, 2011, pp. 189-200.
49. N. Vuokko and P. Kaski. Significance of patterns in time series collections. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM'11)*, Mesa, Arizona, USA, 28-30 April, 2011, SIAM/Omnipress, pp. 676-686.
50. H. Wettig, S. Hiltunen and R. Yangarber. MDL-based modeling of etymological sound change in the Uralic language family. In *Proceedings of the Fourth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2011)*, Helsinki, Finland, 2011.
51. H. Wettig, S. Hiltunen and R. Yangarber. MDL-based models for aligning etymological data. In *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP-2011)*, Hissar, Bulgaria, 2011.
52. K. Zhang and A. Hyvärinen. A general linear non-Gaussian state-space model: Identifiability, identification, and application. In *Proceedings of the Asian Conference on Machine Learning (ACML)*, 2011, pp. 113-128.

Technical reports and other publications

1. T. Ahonen, K. Lemström and S.M. Linkola. Compressing Quantized Tonal Centroid Vectors for Cover Song Identification. In *Proceedings of MIREX*, 2011.
2. T. Elomaa, J. Hollmén and H. Mannila, editors. *Discovery Science — Proceedings of the 14th International Conference (DS 2011)*, October, 2011, Lecture Notes in Computer Science, Vol. 6926, Springer-Verlag.
3. J. Gama, E. Bradley and J. Hollmén, editors. *Advances in Intelligent Data Analysis — Proceedings of the 10th International Symposium on Intelligent Data Analysis (IDA 2011)*, October, 2011, Lecture Notes in Computer Science, Vol. 7014, Springer-Verlag.
4. J. Kivinen, C. Szepesvári, E. Ukkonen and Z.Thomas, editors. *Proceedings of 22nd International Conference on Algorithmic Learning Theory*, 2011, Lecture Notes in Artificial Intelligence, Vol. 6925, Springer-Verlag.
5. L.A. Langohr, V. Podpecan, M. Petek, I. Mozetic and K. Gruden. Subgroup Discovery from Interesting Subgroups. In *Proceedings of Bioinformatics Research and Education Workshop (BREW 2011)*, Estonia, 2011.
6. V. Mäkinen. Algoritmitutkimuksen rooli bioinformatiikassa. *Tietojenkäsittelytiede*, 32, 2011, pp. 10–15.

6. PhD degrees

Members of Algodan obtained 21 PhD degrees, listed below.

2014

1. Galbrun. Esther. Methods for Redescription Mining. *University of Helsinki.*
2. Korhonen, Janne. Graph and Hypergraph Decompositions for Exact Algorithms. *University of Helsinki.*
3. Paalasmaa, Joonas. Monitoring Sleep with Force Sensor Measurement. *University of Helsinki.*

2013

1. Entner, Doris. Causal Structure Learning and Effect Identification in Linear Non-Gaussian Models and Beyond. *University of Helsinki.*
2. Eronen, Lauri. Computational Methods for Augmenting Associations-based Gene Mapping. *University of Helsinki.*
3. Hyttinen, Antti. Discovering Causal Relations in the Presence of Latent Confounders. *University of Helsinki.*
4. Lijffijt, Jeffrey. Computational methods for comparison and exploration of event sequences. *Aalto University.*
5. Luosto, Panu. Normalized Maximum Likelihood Methods for Clustering and Density Estimation. *University of Helsinki.*
6. Timonen, Mika. Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion. *University of Helsinki.*
7. Wettig, Johannes. Probabilistic, Information-Theoretic Models for Etymological Alignment. *University of Helsinki.*

2012

1. Hanhijärvi, Sami. Multiple hypothesis testing in data mining. *Aalto University.*
2. Heinonen, Markus. Computational Methods for Small Molecules. *University of Helsinki.*
3. Parviainen, Pekka. Algorithms for Exact Structure Discovery in Bayesian Networks. *University of Helsinki.*
4. Sirén, Jouni. Compressed Full-Text Indexes for Highly Repetitive Collections. *University of Helsinki.*
5. Vuokko, Niko. Testing the Significance of Patterns in Complex Null Hypotheses. *Aalto University.*
6. Välimäki, Niko. Applications of Compressed Data Structures on Sequences and Structured Data. *University of Helsinki.*
7. Wessman, Jaana. Mixture Model Clustering in the Analysis of Complex Diseases. *University of Helsinki.*
8. Zhou, Fang. Methods for Network Abstraction. *University of Helsinki.*

2011

1. Junttila, Esa. Patterns in Permuted Binary Matrices. *University of Helsinki*.
2. Hintsanen, Petteri. Simulation and Graph Mining Tools for Improving Gene Mapping Efficiency. *University of Helsinki.*
3. Ojala, Markus. Randomization Algorithms for Assessing the Significance of Data Mining Results. *Aalto University.*

APPENDIX 1: Analysis of ALGODAN Publications 2008-2013

The analysis was made by Helsinki University Library, Kumpula Campus Library bibliometrics team in April 2014 using the publication list provided by the research group.

Summary of publication statistics

There are 661 publications listed for 2008-2014. The types and annual statistics are summarized in Table 1.

Table 2: ALGODAN publications by year and type. See explanation below for type classification.

Year	A1	A3-A4	B-E	C	F	G	Total
2008	40	68	10	0	0	6	124
2009	34	46	9	1	0	7	97
2010	34	75	15	0	0	5	129
2011	31	52	6	0	0	3	92
2012	29	51	20	0	1	8	109
2013	32	54	4	0	10	8	108
2014						2	2
Total	200	346	64	1	11	39	661

The publications are classified according to the following scheme:

A1	Articles in refereed scientific journals
A3-A4	Refereed conference articles and articles in edited books
B-E	Technical reports and other publications
C	Books
F	Artistic works
G	Theses

The large share of conference articles and articles in edited books, as shown for A3 -A4, is usual for computer science.

Coverage in bibliometric databases

The coverage of computer science publications in Thomson & Reuters Web of Science is known to be insufficient, and based on this, Web of Science was excluded from this analysis.

A search for all relevant publications (from A1 to C) was made in Elsevier Scopus[3]. The hits are listed in Table 2. Likewise, a search was made with the Publish or Perish tool[4] (based on Google Scholar). The hits are also listed in Table 2.

[3] http://www.scopus.com

[4] http://www.harzing.com/pop.htm

Table 3: Scopus and Publish or Perish hits for ALGODAN publications

Year	Scopus	PoP
2007	1	0
2008	76	109
2009	60	89
2010	99	114
2011	70	82
2012	77	100
2013	52	94
2014	2	2
Total	437	590

Comparing the hits to the data in Table 1, we get the following coverage for ALGODAN publications:

- Scopus 71.5% – 437 out of 611 (artistic works and theses excluded)
- Publish or Perish 90.8% – 590 out of 650 (artistic works excluded)

Publish or Perish (PoP) coverage is good, but identifying the hits was considerably slow and in some cases, it was difficult to identify the hits reliably, so the PoP results should be seen as a good guess.

The coverage in Scopus is better than anticipated, and reflects the efforts by Elsevier to increase the number of conference publications in Scopus. The publication types (for more than 1 hits) are listed in Table 3. The most popular publication channel was LNCS as seen from Table 4.

Table 4: Scopus publication types with > 1 hits

Conference Paper	256
Article	163
Review	7
Book Chapter	4
Letter	3

Table 5: Publication channels with >4 hits according to Scopus

SOURCE TITLE	# of titles
Lecture Notes in Computer Science	125
Journal of Machine Learning Research	13
Information Processing Letters	12
Proceedings IEEE International Conference on Data Mining	9
ACM International Conference Proceeding Series	7
BMC Bioinformatics	6
Proceedings of the ACM SIGKDD	6
Neurocomputing	5
Bioinformatics	5
Plos One	5

Scopus seems to be a little inconsistent with subject areas. A few publication series that probably should be classified under computer science are classified e.g. under mathematics only. A rough view of subject area provided by Scopus looks like in Figure 1.

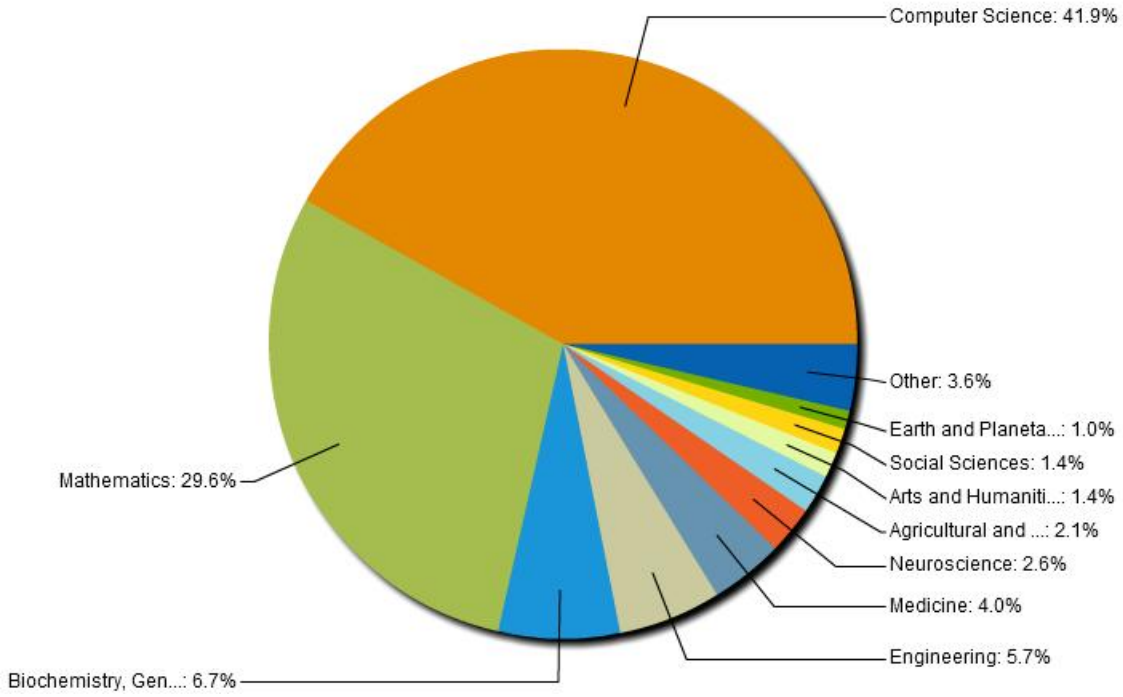

Figure 1: ALGODAN subject areas in Scopus

Citation analysis with Scopus

The ALGODAN publications found in Scopus received 2673 citations by beginning of April 2014. A more detailed view for each year is seen in Table 5.

Table 6: Citations in Scopus by year

Year	2008	2009	2010	2011	2012	2013	2014	Total
2007	1	0	0	0	0	0	0	1
2008	22	96	162	158	180	133	26	777
2009		37	152	167	159	146	27	688
2010		2	53	243	263	266	39	866
2011		1	0	8	69	92	13	183
2012					26	80	14	120
2013						30	8	38
2014							0	0
Total	23	136	367	576	697	747	127	2673

As University of Helsinki does not have access to Scopus raw data, we could only do a partial analysis of highly cited (top 10%) publications. The annual volume of Computer Science publications in Scopus is around 300,000 while Scopus only allows downloads of max 20,000 items. However, it was possible to

extrapolate the citation numbers to get the cut-off value for the top 10% for each year for computer science. As for publications that were classified under mathematics, the annual volume was low enough and no extrapolation was needed.

For biosciences and medicine, it was not possible to estimate a 10% cut-off value due to high publication volumes. Also, the citation rates for these subject areas differ strongly from one subfield to another.

Table 6 lists the number of highly cited publications for the two dominant subject areas. The mathematics values cover only those publications not listed under computer science.

Table 7: Number of highly cited publications

		Computer science		Mathematics		
year	Scopus hits	*10% cutoff*	# of10% items	*10% cutoff*	# of 10% items	# of 10% items
2008	76	*12*	13	*14*	3	16
2009	60	*9*	12	*11*	1	13
2010	99	*6*	17	*9*	0	17
2011	70	*6*	7	*6*	1	8
2012	77	*3*	6	*3*	1	7
2013	52	*1*	9	*1*	2	11

A list of higly cited publications can be found in Appendix 1.

Citation analysis with Publish or Perish

Citations from Publish or Perish are listed in Table 7. The quality of the citations has not been controlled in any way.

Table 8: Citations by year and type according to PoP

Year	A1	A3-A4	B-E	C	Total
2008	1111	752	23	0	1886
2009	1006	532	9	208	1755
2010	1102	386	22	0	1510
2011	270	166	0	0	436
2012	175	136	25	0	336
2013	149	65	29	0	243
Total	3813	2037	108	208	6166

Co-operation and co-authorship

Table 7 lists the countries and Table 8 the institutional affiliations of the contributing authors.

Table 9: Countries of ALGODAN authors according to Scopus (>4)

COUNTRY	#
Finland	425
United States	79
United Kingdom	28
Germany	26
Italy	20
Sweden	19
Japan	18
Spain	16
France	15
Canada	10
Chile	10

COUNTRY	#
Greece	10
Poland	10
Netherlands	9
Denmark	9
Australia	9
Belgium	6
Israel	5
New Zealand	5
Norway	5
Switzerland	5

Table 10: Institutional affiliations of ALGODAN authors according to Scopus (>4)

Affiliation	#
University of Helsinki	342
Aalto University	146
Helsinki Institute for Information Technology	23
Stony Brook University State University of New York	17
Universidad de Chile	12
Helsinki University Central Hospital	10
Osaka University	9
University of Texas at Arlington	8
University of Helsinki Institute of Biotechnology	8
Massachusetts Institute of Technology	8
Lunds Universitet	8
VTT Technical Research Centre of Finland	8
Yahoo Research Barcelona	8
University of Athens	7
King's College London	7
IT-Universitetet i København	7
University of Arizona	7
Carnegie Mellon University	6
Jozef Stefan Institute	5
European Commission Joint Research Centre, Ispra	5