Algorithmic Data Analysis (Algodan) Centre-of-Excellence

Biennial report 2010-2011

May 15, 2012

Esko Ukkonen, Pirjo Moen, eds.

Contents

Preface

The Finnish Centre of Excellence in Algorithmic Data Analysis Research (Algodan) is now in its fifth year of operations. This report describes the activities of the Centre in 2010 and 2011, and also looks forward into the future years. The report is organised according to the teams and their research groups. The groups present their members, mission and main results as well as future plans, cooperation and societal impact. Each group description is followed by a short list of its most important publications. The complete lists of publications and PhD degrees are presented at the end of the report as well as the funding of the centre.

Helsinki, May 15, 2012

Esko Ukkonen

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

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 31 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.

Reports from the Teams and their Groups

Team Data Mining: Theory and Applications

Data mining – theory and applications

Members

- Kai Puolamäki, PhD, Group leader
- Heikki Mannila, Professor, on leave of absence
- Panagiotis Papetrou, PhD, Postdoctoral researcher
- Jefrey Lijffijt, Doctoral student
- Aleksi Kallio, Doctoral student, part-time
- Recent alumni:
    - Niko Vuokko, Doctoral student until 2011, doctoral defense on February 2012.
    - Markus Ojala, Doctoral student until 2011, doctoral defense on November 2011.
    - Esa Junttila, Doctoral student until 2011, doctoral defense on August 2011.
    - Sami Hanhijärvi, Doctoral student until 2010, doctoral defense on May 2012.

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.

Research activities

S - sequence analysis

The highlights of the group's research activities in 2011 include our continuing work on randomization methods and statistical significance testing in data analysis. We have applied the randomization and statistical testing methodology especially in the analysis of sequences.

As an example of the analysis of (word) sequences, we have shown that even the simplest statistics such as the word frequency in the text may be uninformative, if typical assumptions such as bag of words are used [6]. This is part of our focus on computational linguistics in collaboration with the group of Prof Terttu Nevalainen (also see [7]). On bioinformatics and sequences, Kallio et al. [5] apply randomization methods to assess the significance of gene periodicity results. Papapetrou et al. [9] study subsequence matching in time series databases.

D - discovery of hidden structure in high-dimensional data

The group has continued its collaboration with the ecologists. The group's objective has been to introduce novel computational methods and to solve algorithmic problems that are related to biotic interactions and climate. As an example, we have on one hand been able to estimate the precipitation based on morphology of (fossil) tooth [1, 2]. On the other hand, we have studied of how to use correlations to answer questions from ecologican spatiotemporal presence-absence data [4].

F - Foundations of algorithmic data analysis

Ojala et al. [8] study how to assess performance of a classifier using randomization. The topic of the Esa Junttila's doctoral thesis was nestedness, a new computational concept on binary data [3].

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. The publication record and the collaboration network of the group speak for itself.

The alumni of the group have moved to responsible positions in academia and industry. Prof Heikki Mannila, who established the group, was nominated in 2009 as the Vice President of then newly established Aalto University. Beginning year 2012, Prof Mannila is serving as the President of the Academy of Finland.

The group is not currently actively hiring new members. The group continues to function, but the activities will be tuned down as its members move forward in their careers. The definite strength of the group has been the focus on quality, both in recruitment and selection of research topics, collaborators, and publication venues. The focus of quality will be even more important now as the quantity is not increasing.

Societal, economic, and technical impact

The main societal impact of the work is in the use of the methods in other sciences. As a recent example, our methodological work on ecology has been received very well, evident for example from the citation counts (e.g., Eronen et al. [1, 2] are already in the 10+ citations range, according to Google Scholar).

The alumni of the group have positioned themselves well in industry and academia, Antti Ukkonen and Nikolaj Tatti being currently on a postdoctoral period abroad. Prof Mannila is the President of the Academy of Finland.

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.

Publications

1. J. T. Eronen, K. Puolamäki, L. Liu, K. Lintulaakso, J. Damuth, C. Janis, and M. Fortelius. Precipitation and large herbivorous mammals, part I: Estimates from present-day communities. Evolutionary Ecology Research, 12(2), 2010, pp. 217–233.
2. J. T. Eronen, K. Puolamäki, L. Liu, K. Lintulaakso, J. Damuth, C. Janis, and M. Fortelius. Precipitation and large herbivorous mammals, part II: Application to fossil data. Evolutionary Ecology Research, 12(2), 2010, pp. 235–248.
3. Gemma Garriga, Esa Junttila, and Heikki Mannila. Banded structure in binary matrices. Knowledge and Information Systems, 28(1), 2011, pp. 197-226.
4. Aleksi Kallio, Kai Puolamäki, Mikael Fortelius, and Heikki Mannila. Correlations and co-occurrences of taxa: the role of temporal, geographic and taxonomic restrictions. Palaeontologia Electronica, 14(1), 2011.

5. Aleksi Kallio, Niko Vuokko, Markus Ojala, Niina Haiminen, and Heikki Mannila. Randomization Techniques for Assessing the Significance of Gene Periodicity Results. BMC Bioinformatics, 12, 2011, p. 330.
6. Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In Proc ECML PKDD 2011.
7. Terttu Nevalainen, Helena Raumolin-Brunberg, and Heikki 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.
8. Markus Ojala and Gemma Garriga. Permutation Tests for Studying Classifier Performance. Journal of Machine Learning Research, 11, 2010, pp. 1833-1863.
9. Panagiotis Papapetrou, Vassilis Athitsos, Michalis Potamias, George Kollios, and Dimitrios Gunopulos. Embedding-based Subsequence Matching in Time Series Databases. In ACM TODS, 36(3), 2011, p. 17.

Parsimonious Modelling Group

Members

- Jaakko Hollmén, Chief Research Scientist, Group leader
- Mika Sulkava, Academy Postdoctoral Researcher, member until 2011
- Miguel Angel Prada, PhD, member until 2010
- Mikko Korpela, M.Sc.(Tech.), Doctoral student
- Janne Toivola, M.Sc.(Tech.), Doctoral student
- Prem Raj Adhikari, M.Sc.(Tech.), Doctoral student
- Olli-Pekka Rinta-Koski, M.Sc.(Tech.), Doctoral student
- Andres Sanz Garcia, University of La Rioja, Spain, visitor 3 months in 2011
- Serafin Alonso Castro, University of Leon, Spain, visitor 3 months in 2011
- Antonio Morán, University of Leon, Spain, visitor 3 months in 2011

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]. 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.

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].

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, climate) as well as the man-made environment.

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]. 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 during 2011. Good balance between applications and methodologies has been achieved.

Future plans

Research will be continued in all areas. Multiresolution data analysis is now investigated in both cancer genomics and environmental informatics, which could bring synergetic effects.

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 and string matching

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.

Combinatorics, Algebra, and Computing (CO-ALCO)

Members

- Mikko Koivisto, Academy Research Fellow, Co-leader
- Petteri Kaski, Academy Research Fellow (9/2011-), Professor (1/2012-), Co-leader
- Pekka Parviainen, Doctoral student (PhD March 2012)
- Janne Korhonen, Doctoral student
- Juho-Kustaa Kangas, MSc student (8/2011-)

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

Concerning combinatorial structures in binary data matrices we have studied the algorithmic and modeling aspects of what we call segmented nestedness [5].

We have continued our research on algorithmic foundations of learning graphical models. We extended our partial order methodology for trading time for space [7] to a Bayesian approach to structure discovery in Bayesian networks [8]; for implications towards applications, see the report of the Phenomics group. For parametrized variants of so-called polytrees, we showed tractability and intractability results [3].

F - Foundations of algorithmic data analysis

We generalized our earlier space-time tradeoff approach in two directions: the original technical innovation was formulated as a combinatorial problem of “covering” all linear orders on a finite set by a relatively small number of "thin" partial orders; and, the approach was shown to work for a large class of permutation problems, including, e.g., the traveling salesman problem. The work was presented at SODA’10 [8].

Motivated the power of fast zeta and Möbius transforms in algorithm design, as proved in our previous research, we have continued studying combinatorially flavored variants of these transforms. We discovered a split-transform technique that yields significant space savings at only negligible increase in the runtime [1]. Another highlight concerns the algorithmics of Möbius inversion on finite lattices: 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 algebra; this work was presented at SODA’12 [2].

Other highlights are the nonconstructive enumeration of the main classes of Latin squares of order 11 [4] and our study of the relative power of OR and SUM circuits for the so-called ensemble problem [6].

Future plans

The structure of the group is in transition. Petteri Kaski is about to build his own group within Algodan. Pekka Parviainen is going to start a post-doc period elsewhere in June 2012. What remains will be merged with the Phenomics group into a new group, with a mission similar to that of the present CO-ALCO but emphasizing the role of so-called sum-product problems. The new groups of Kaski and Koivisto will work in close collaboration; below we discuss some shared future research plans.

Our previous works have brought up open questions in extremal combinatorics, answering which would yield better complexity estimates and, potentially, algorithm designs; our current interests are especially in the number of connected sets in bounded degree graphs and in the tradeoff of the number of linear extensions and downsets of partial orders. Work on classification problems will continue. Computer-aided analyses are central in these studies.

Concerning the development of novel algorithmic techniques, we will turn an increased attention to various tradeoffs in resources and objectives: e.g., time versus space, time versus success probability, time versus parallelization. These perspectives are somewhat fresh and, in addition to recently emerged theoretical interests, are expected to extend the scope of our techniques on practical algorithmics.

We will continue studies in algorithmic methods in data analysis, focusing on the challenge of discovering Bayesian networks from observational data. Compared to our previous works, we will put more emphasis on principled approximative methods and restricted model classes.

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 societal impact in the long run, say within the next fifty years.

To mention a specific example of impact in the research community, our fast subset convolution framework (Björklund et al., STOC 2007) has attracted a fair number of citations (~110 on Google Scholar by April 2012), and our results on exponential algorithms amount to about 35 pages (20 %) of the book “Exact Exponential Algorithms” by Fomin and Kratsch (Springer, 2010). We expect our future results have similar impact.

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. 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.
- 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. Andreas Björklund, Thore Husfeldt, Petteri Kaski, and Mikko Koivisto. Covering and packing in linear space. Information Processing Letters 111 (2011) 1033-1036.
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.
3. Serge Gaspers, Mikko Koivisto, Matthieu Liedloff, Sebastian Ordyniak, and Stefan Szeider. On finding optimal polytrees. 26th Conference on Artificial Intelligence (AAAI 2012), to appear.
4. Alexander Hulpke, Petteri Kaski, and Patric R. J. Östergård. The number of Latin squares of order 11, Mathematics of Computation 80 (2011) 1197-1219.
5. Esa Junttila and Petteri Kaski. Segmented nestedness in binary data. 11th SIAM International Conference on Data Mining (SDM 2011), pp. 235-246, SIAM / Omnipress, 2011.
6. 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), to appear.
7. Mikko Koivisto and Pekka Parviainen. A space-time tradeoff for permutation problems. 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2010), pp. 484-492, SIAM, 2010.
8. Pekka Parviainen and Mikko Koivisto. Bayesian structure discovery in Bayesian networks with less space. 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010), Vol. 9 of JMLR: W&CP 9, pp. 589-596.

Phenomics Group

Members

- Heikki Mannila, Professor, Group leader (-2/2012)
- Mikko Koivisto, Academy Research Fellow, Co-leader
- Stefan Schönauer, Postdoctoral researcher (- 12/2010)
- Jussi Kollin, Doctoral student (-10/2010, PhD 10/2010)
- Pekka Parviainen, Doctoral student (PhD 3/2012)
- Jaana Wessman, Doctoral student (-12/2011, PhD 4/2012)
- Teppo Niinimäki, Doctoral student

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

The project on computational methods for the discovery of large-scale chromosomal rearrangements (such as deletion and inversion polymorphisms) in SNP data was completed. Our methods, while sound and computationally efficient, turned out to have an equal or lower statistical power compared to similar methods that had been developed independently by Benjamin Raphael's group at Brown University. The results are published in the PhD thesis of Jussi Kollin (2010).

A related high-risk project on the discovery of inter-chromosomal associations was completed. By extensive computational analyses of several SNP datasets, we located a number of marker pairs that showed statistically significant associations much beyond what could be expected by chance. By a series of careful further analyses the findings were, however, classified to previously known or unknown database errors and associations that could be explained by the fact that the genotype data were collected non-uniformly, i.e., using different platforms for different data subsets. The results are published in the MSc thesis of Teppo Niinimäki (2010).

We have also 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 will be published in PLos ONE [2]; 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 and ECML-PKDD'11 conferences, respectively [1, 2].

Future plans

The structure of the group is in transition, as many members have left the group. We plan to merge the remaining group with a part of the CO-ALCO group into a new group within Algodan; see a related description about CO-ALCO.

How to best discover and analyze causal and statistical dependences between hundreds of phenotypes (and genotypes), especially when the number of observed subjects is only some thousands? We plan to continue approaching this question from the following two main angles. First, we continue the work in a direction opened up by our new MCMC sampling scheme [1]. Second, we study to what extent it is possible to discover the dependency network by local analyses, that is, by discovering the neighborhood (Markov blanket) of each variable of interest by a method that ignores distant dependencies that, in a sense, are outside the neighborhood. Our methods will be compared to existing methods using both simulated data and the NFBC66 data; the results will be constantly discussed with domain experts.

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).

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. 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 Press, 2011.
2. 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.
3. 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, in press.

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

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

The so-called position weight matrix (PWMs) is the standard model for the binding affinity distribution between a transcription factor (TF) and DNA. PWMs are used for finding putative binding sites of TFs in DNA by searching through genomic DNA sequences. Our earlier work on PWM search, applying sophisticated techniques from string matching algorithms to achieve order-of-magnitude speed-up of the performance, was published [7]. The binding sites often have gapped structure; search algorithms for this special case and related computational complexity results were reported in [1, 8].

High-quality PWM models for TFs are an essential component of the research of gene regulatory systems. 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. To analyze this exceptional data, we have developed a novel learning algorithm for PWMs [2, 5]. Moreover, we have developed new models and learning algorithms for binding sites of TF complexes that consist of more than only one factor (dimers, trimers) [11]. The long-term goal is a Markov chain model that would give essentially more accurate predictions of binding sites than current models whose false-positive rate is still very high.

The group participates in Professor Ilkka Hanski's (Metapopulation Research Unit, Department of Ecology and Evolutionary Biology) major project of de novo sequencing of the entire genome of the Glanville fritillary butterfly (Melitaea cinxia); this is the first eukaryote sequencing project in Finland. We have developed the data analysis pipeline for high-throughput sequencing, with novel algorithms for error correction [6, 12] and scaffolding [9]. The genome reconstruction is expected to be completed in the first half of 2012.

We have produced several results in basic research in algorithms for sequences. A useful observation on using string matching technique in point pattern matching was published [3]. We proved a practical result about the optimality of a class of string searching algorithms in the average case [13]. We have also introduced an alphabet sampling technique to speed up string searching algorithms [14].

D - Discovery of hidden structure in high-dimensional data

In a joint project with European Bioinformatics Institute (Alvis Brazma, Margus Lukk), we analyzed a globally collected dataset of gene expression in humans (Affymetrix U133A microarray data on 5372 samples from 206 different studies generated in 163 different laboratories). By applying various clustering and dimension reduction approaches to this very high dimensional (18609 dimensions) data we found quite unexpectedly a 'continent' structure, i.e., a small number of distinct expression profile classes, having quite natural biological interpretation (http://www.genomeweb.com/informatics/ebi-helsinki-team-integrates-array-data-thousands-samples-map-global-expression-) [4].

In a joint project with the Division of Astronomy of the university (Professor Kalevi Mattila & 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. As the first reported result, 137 previously unknown cluster candidates and 30 previously unknown sites of star formation were found [10].

Future plans

We are currently working on the following projects:

1. 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 tools for modeling gene regulatory relations and disorders in them.
2. The current genome project of the Glanville fritillary butterfly is about to be finished. Next we plan to join, in similar role, the birch tree genome project of Professor Jaakko Kangasjärvi. Based on our experience on two major genome reconstruction projects, we may develop our sequencing software into a publicly available package.
3. We will continue the analysis of the Infrared Deep Sky Survey data by developing alternative search algorithms that should make the search more accurate and automatic by eliminating more hand work.

Another planned project is on computational analysis of micro-RNA profiles for guided differentiation of functional retinal pigment epithelium, with Professor Arto Urtti (Centre for Drug Research). Here our role is to model the guided differentiation process.

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), Kalevi Mattila (Astronomy), Jussi Taipale (Biomedicum & Karolinska Institutet, Sweden),Jaakko Kangasjärvi (Dept of Biology) and Arto Urtti (Drug Research).

Other national cooperation: VTT Biotechnology, Professor Merja Penttilä (metabolic modeling).

International cooperation:

1. Dr Alvis Brazma, European Bioinformatics Institute, UK; analysis of gene expression data; joint publications and researcher education (Margus Lukk);
2. Professor Alberto Apostolico & Dr. Cinzia Pizzi, University of Padua, Italy and Georgia Tech, USA; analysis of motifs in strings; joint publications.

Societal, economical and technical impact

Some of our new algorithms (for the genome assembly [6, 9, 12] or for learning PWMs [2], for example) have potential to become into wide use as they offer improved performance. On application side, for example the findings on the clustered structure of the gene expression space [4] or new PWM models that are more accurate than the earlier ones [11] have potential for a significant impact in their fields. It is too early to evaluate the impact from citations but it seems that [2, 4, 5] are attracting a rapidly growing number of them.

Selected publications

1. M. Michael, F. Nicolas & E. Ukkonen.On the complexity of finding gapped motifs. Journal of Discrete Algorithms 8 (2010), 131–142.
2. 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.
3. E. Ukkonen. Geometric Point Pattern Matching in the Knuth-Morris-Pratt Way. Journal of Universal Computer Science 16, 14 (2010), 1902–1911.
4. M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen & Alvis Brazma. A global map of human gene expression. Nature Biotechnology 28, 4 (April 2010), 322–324.
5. 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.
6. Leena Salmela. Correction of sequencing errors in a mixed set of reads. Bioinformatics 26 (10): 1284-1290 (2010).
7. C. 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.
8. Apostolico, C. Pizzi & E. Ukkonen. Efficient Algorithms for the Discovery of Gapped Factors. Algorithms for Molecular Biology 6:5 (2011), 10 pages.
9. L. Salmela, V. Mäkinen, N. Välimäki, J. Ylinen & E. Ukkonen. Fast scaffolding with small independent mixed integer programs. Bioinformatics 27, 23 (2011), 3259–3265.
10. O. Solin, E. Ukkonen & L. Haikala. Mining the UKIDSS GPS: star formation and embedded clusters. Astronomy & Astrophysics, 2012 (in press).
11. A. Jolma, J.Yan, Th. Whithington, J. Toivonen et al. Binding specifities of human transcription factors. Manuscript, April 2012.

12. L. Salmela & J. Schröder. Correcting errors in short reads by multiple alignments. Bioinformatics 27(11):1455-1461 (2011).
13. L. Salmela. Average complexity of backward q-gram string matching algorithms. Information Processing Letters, Volume 112(11):433-437 (2012).
14. F. Claude, G. Navarro, H. Peltola, L. Salmela & J. Tarhio. String matching with alphabet sampling. Journal of Discrete Algorithms 11:37-50 (2012).

Succinct Data Structures (SuDS)

Members

- Veli Mäkinen, Professor, Group leader
- Niko Välimäki, Doctoral student
- Jouni Sirén, Doctoral student
- Riku Katainen, Research Assistant (-8/2010)
- Juha Karjalainen, Research Assistant (9/2010-12/2010)
- 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.

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

Societal, economical and technical impact

The publications are still quite fresh to have gathered lots of citations. The earliest [1] has 18 citations. The group has from year 2007 one publication with 333 citations and another from the same year with 193 citations, so it is likely that some of the new developments reach similar levels. There has been a big paradigm shift towards bioinformatics, and it is quite likely that the citation rates go down until there is a strong enough result that gathers attention within the new audience. However, we are confident that our new developments in [6] will be very useful in the variation calling application in our new context of cancer genetics research. Also the work on XML indexing should have important role in future infrastructure for data storage and retrieval in specific domains.

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 on Strings

Members

- Juha Kärkkäinen, University Researcher, Group leader
- Simon Puglisi, Postdoctoral researcher
- Dominik Kempa, Doctoral student
- Pekka Mikkola, MSc student

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

The Burrows-Wheeler transform (BWT) has a central role in many data compressors including the popular bzip2 program. The BWT is the bottleneck in the compression phase and the inverse BWT in the decompression phase. We have developed several new algorithms for the inverse BWT [1,5] including an algorithm that is consistently about four times faster than the previously fastest one. We have also improved the speed of both compression and decompression by performing grammar compression before BWT [8]. These improvements are a part of an experimental, open source compressor available at https://github.com/pjmikkol/bwtc.

The BWT is also used in compressed text indexes that not only compress the text but support fast pattern matching queries. We have developed a new technique called fixed-block compression boosting [3] that improves BWT-based compressed text indexes in three ways: better compression, faster queries and simpler implementation.

The above practical results above are based on better understanding of the combinatorial properties of the BWT including some new results discovered and proved by us [3,5].

On a (so far) more theoretical level, we have studied other types of compressed text indexes [4], and the problem of performing a large number of pattern matching queries simultaneously [6,7]. We have also developed a data structure that stores a text in compressed form and can quickly count the number of distinct symbol in a query region [2]. Besides compressing the text better than previous indexes, it is the first such data structure where the query time does not depend on the length of the text, only on the size of the query region.

Future plans

The BWT transforms the text but does not compress it; this is done by the entropy compression stage that follows BWT. This stage is the next focus of our research both in our experimental compressor and in compressed text indexes. The techniques currently used in compressors are quite different from the ones in compressed indexes, but we believe that studying both together leads to better understanding and better algorithms for both. We are also directing our research towards other types of compressors and compressed indexes.

We are continuing research on indexed multiple pattern matching including implementation and experimentation.

Cooperation

Cooperation within Algodan

We have worked on indexed multiple pattern matching [6, 7] with the groups of Veli Mäkinen and Esko Ukkonen, and on index construction of huge genomic databases with the group of Veli Mäkinen.

National cooperation

National cooperation: Professor Jorma Tarhio, Travis Gagie, and Kalle Karhu, Aalto University; compressed indexes, indexed multiple pattern matching; joint publications.

International cooperation

International cooperation: Simon Gog, University of Ulm, Germany; indexed multiple pattern matching, joint publication, research visit. Paweł Gawrychowski, Max-Planc-Institut für Informatik, Saarbrücken, Germany and Yakov Nekrich, University of Bonn, Germany; compressed indexes; joint publication.

Societal, economical and technical impact

The fast and practical algorithms and techniques for fundamental problems developed by the group have the potential for being included in many applications. In particular, we believe that the fast inverse BWT algorithm [5] and the fixed block compression boosting technique [3] will become standard practices as they are simple and offer significant practical benefits. With increasing amounts of data and wide use compression in data storage and communication, better algorithms and techniques can lead to economic benefits too.

Selected publications

1. Juha Kärkkäinen, Simon Puglisi. Medium-Space Algorithms for Inverse BWT. In Proc. 18th European Symposium on Algorithms (ESA 2010), Springer, 2010, pp. 451-462.
2. Juha Kärkkäinen, Travis Gagie. Counting Colours in Compressed Strings. In Proc. 22nd Symposium on Combinatorial Pattern Matching (CPM 2011), Springer, 2011, pp. 197-207.
3. 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.
4. Travis Gagie, Paweł Gawrychowski, Juha Kärkkäinen, Yakov Nekrich and Simon J. Puglisi. A Faster Grammar-Based Self-Index. In Proc. 6th Conference on Language and Automata Theory and Applications (LATA 2012), Springer 2012, pp. 240-251.
5. Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. Slashing the Time for BWT Inversion. In Proc. 2012 Data Compression Conference (DCC), IEEE Computer Society 2012, pp. 99-108.
6. Travis Gagie, Kalle Karhu, Juha Kärkkäinen, Veli Mäkinen, Leena Salmela, Jorma Tarhio. Indexed Multi-Pattern Matching. In Proc. 10th Latin American Theoretical Informatics Symposium (LATIN 2012), Springer 2012, pp. 399-407.
7. Simon Gog, Kalle Karhu, Juha Kärkkäinen, Veli Mäkinen, Niko Välimäki. Multi-Pattern Matching with Bidirectional Indexes. To appear in 18th International Computing and Combinatorics Conference (COCOON 2012).

8. Juha Kärkkäinen, Pekka Mikkola, Dominik Kempa. Grammar Precompression Speeds Up Burrows-Wheeler Compression. Submitted to a conference, 2012.

Computational Geometry

Members

- Valentin Polishchuk, Academy postdoctoral researcher, Group leader
- Mikko Nikkilä, MSc student
- Mikko Sysikaski, MSc student
- Juha-Antti Isojärvi, MSc student
- Sylvester David Eriksson-Bique, Doctoral student

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

Our research proceeds along 4 major directions:

1. Multi-agent motion coordination is the objective in a number of applications: swarm robotics, battlefield operations, ship routing, evacuation planning, video games and simulators. Our particular motivation for studying route planning for multiple objects comes from air traffic industry needs where the task is to plan and monitor motion of a multitude of entities (aircraft) amidst dynamic and stochastic obstacles (hazardous weather cells).
   We developed advanced geometric techniques for computing multiple paths for different-type agents amidst various constraints [1,2].
2. Map labeling deals with visual delivery of information present on a map. We considered issues arising due to the dynamic nature of interactive (e.g., online) maps – how should the labels behave as the user zooms and pans the map [4].
3. Shape approximation allows one to reconstruct the shape of an unknown geometric object from a set of (often noisy) samples, as well as to simplify the description of a given object. We designed improved algorithms for finding low-complexity approximations of paths and polygons, both in two and three dimensions [5,6].
4. Sensor network is a source of a bulk of measurements taken by the sensors over time. We addressed a variety of computational-geometry questions that arise from gathering and processing sensor-network data [3].

Highlights

1. Multicommodity network flows are a classical subject in optimization; we extended the study of multicommodity flows to geometric domains. The results were presented both in a purely theoretical journal [1] and in the top application-oriented outlet [2].

2. Localization, clustering, and activity scheduling are fundamental problems in wireless sensor networks. We developed a unified approach to understanding spatial dependency and correlation between the measurements, sensor localization, sensor suspension and system lifetime maximization. We addressed the questions both in centralized and distributed settings [3].

Future plans

Our plan is to to continue research in geometric methods for a variety of applications. On the motion planning frontier, to investigate further fundamental geometric problems of planning paths under a multitude of constraints and requirements, in particular, study different-homotopy paths and curvature-constrained motion. For the shape approximation, design simplified and improved ``coresets'' – small representative subsets of point clouds; automate shape reconstruction approaches to seismic data analysis. For sensor networks, design optimal protocols for data collection, address the questions of data security and transmission interference.

Cooperation

Cooperation within the universities

- Jukka Suomela. HIIT. Local algorithms. Joint publications.

International cooperation:

- Stony Brook University, USA. Computational geometry. Long-standing collaboration with mutual visits and a dozen of co-authored articles especially on motion coordination; research visits.
- Metron Aviation Inc., USA. Path planning for air traffic management. Long-standing collaboration; co-authored articles.
- Karlsruhe Institute of Technology, Germany. GIS. 1 research article.
- Institute for Dynamics of Geospheres, Russia. Seismological data analysis. Research visit from them.
- University of Arizona, USA. Data processing in sensor networks. Long-standing collaboration; coauthored articles.
- University of British Columbia, Canada. Motion planning. Mutual research visits.

Participation in European research networks

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

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.

Easy-to-follow maps can be deployed both in static settings (bus stops, tourist booklets) and as interactive tools (integrated, e.g., into route planners); map readability is also important when using a mobile device. 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.

Selected publications:

1. J. Kim, J. S. B. Mitchell, V. Polishchuk, S. Yang, J. Zou. Routing Multi-Class Traffic Flows in the Plane. CGTA, 45(3):99-114, 2012.
2. S. Yang, J. S. B. Mitchell, J. Krozel, V. Polishchuk, J. Kim, J. Zou. Flexible Airlane Generation to Maximize Flow Under Hard and Soft Constraints. Air Traffic Control Quarterly 19(3):1-26, 2011.
3. P. Agarwal, A. Efrat, C. Gniady, J. S. B. Mitchell, V. Polishchuk, G. Sabhnani. Distributed Localization and Clustering Using Data Correlation and the Occam's Razor Principle. DCOSS'11
4. M. Nöllenburg, V. Polishchuk, M. Sysikaski. Dynamic One-Sided Boundary Labeling. ACM SIGSPATIAL GIS'10.
5. V. Polishchuk, M. Sysikaski. Faster algorithms for minimum-link paths with restricted orientations. WADS'11 .
6. E. Arkin, A. Efrat, G. Hart, I. Kostitsyna, A. Kröller, J. S. B. Mitchell, V. Polishchuk. Scandinavian Thins on Top of Cake: on the Smallest One-Size-Fits-All Box. FUN'12.

C-BRAHMS Group

Members

- Kjell Lemström, PhD, 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)

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].

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].

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. The book that will be one of the first textbooks in the area, will be published by the Oxford University Press in 2013.

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.

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.

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
- Costas Iliopoulos, Maxime Crochemore, King's College, 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.

Computational Linguistics Group

Members

- Roman Yangarber, Principal Investigator, Group leader
- Mian Du, MSc student
- Peter von Etter, MSc, Researcher
- Silja Huttunen, Doctoral student (Linguistics) (until 2011)
- Arto Vihavainen, MSc student (until 2011)
- Suvi Hiltunen, MSc student (until 2011)
- Mikhail Novikov, MSc student (until 2011)
- Lidia Pivovarova, Doctoral student (2011--2013, from St Petersburg State University)
- Hannes Wettig, Doctoral student
- Javad Nouri, MSc student
- Guowei Lv, MSc student
- Matthew Pierce, MSc student

Mission of the group

The group works on problems in analysing 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 three domains of news: epidemiological surveillance, business intelligence, 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. One benefit of engaging end-users is that they provide high-quality 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 the Finnish language in the context of languages genetically related to it – viz., the Uralic language family.

This involves using databases for these languages and developing computational approaches to study of language relationships (currently applied to the Uralic and Turkic families, but applicable generally as well) and language structure. 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 language 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 the 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 research objectives are:

1. investigate relationships between Finnish, the Uralic family, and its neighbors;
2. collect and organise resources of relevant data;
3. develop novel computational methods for discovering and investigating the relationships;
4. apply the methods in a wider setting to investigate deeper connections beyond the Uralic family, and to the study of other language families.

We do not aim to replace the human computational linguist by a model, but 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 organizations – KOTUS and the Russian Academy of Sciences.

Future plans

In PULS, one new direction the group is exploring is the acquisition of different types of knowledge bases in parallel, where they provide mutual constraints to maintain high precision. Another important new direction is extending the linguistic coverage of PULS; most of the work to date has focused 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 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.

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

a) to extend the methods for alignment of observed data to perform reconstruction of the corresponding forms in the unobserved, ancestral languages.
b) to explore more distant and speculative connections reaching further beyond the Uralic family,
c) to apply the methodology to other language families, especially the less-studied families, to help obtain novel results.

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 closely 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.

Cooperation among Universities

The PULS project collaborates with the group of Timo Honkela at the Aalto University, through the joint ContentFactory project.

International cooperation

PULS and Etymon have strong 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 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, radio-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 World Health Organization, as well as several large national health ministries. As mentioned earlier, 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.

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 Organization). 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 Nokia, Esmerk Oy.

The results of Etymon will provide insights into the structure and development of the Uralic language family, as well as other language families. These relationships have been studied for two centuries using manual methods; computational methods for analysing this type of data are only beginning to emerge. The results will have wide-ranging implications for the understanding of the common origins of language.

Publications

1. Hannes Wettig, Kirill Reshetnikov and Roman Yangarber. Using Context and Phonetic Features in Models of Etymological Sound Change. In EACL 2012: Workshop on Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources (2012) Avignon, France.
2. 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.
3. 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
4. Mian Du, Peter von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva, Roman Yangarber. Building support tools for Russian-language information extraction. BSNLP-2011: Balto-Slavonic Natural Language Processing (2011) Plzeň, Czech Republic.
5. 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).
6. 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.
7. 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.
8. Hannes Wettig, Suvi Hiltunen, Roman Yangarber. Hidden Markov models for induction of morphological structure of natural language. WITMSE-2010: Workshop on Information Theoretic Methods in Science and Engineering (2010) Tampere, Finland.

Team Link and Pattern Discovery

Discovery Group: Data Mining and Computational Creativity

Members

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

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 recently also on their use in creative systems. Application areas range from link discovery in bioinformatics to computational generation of poetry.

Research activities

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

We have recently developed a range of novel methods to simplify large networks into simpler ones or for extracting relevant information from them [6-9]. These methods allow more efficient and user-friendly analysis of social networks, biological networks, etc.

The more applied line of this research has produced Biomine, a search engine prototype that integrates and indexes data from several publicly available biological databases [2]. Biomine presents the data as a weighted graph, and its query tools aid explorative discovery of non-trivial connections between biological entities, such as genes and phenotypes. See http://biomine.cs.helsinki.fi/

A new focus area is computational creativity, interesting on its own right but also as an application area for data mining methods. In 2011, we initiated work on computational poetry. We are developing novel methods that minimize the need for manually coded or language specific knowledge [5]. A new postdoc, A. Valitutti, also works on computational humour [11].

We have also continued international collaboration in data mining research, on bioinformatics [3] and database integrity constraints [1]. We have also applied data mining in learning technology [4].

Future plans

The group will continue its work on data mining, especially on problems related to information networks. The new research area of computational creativity will be nurtured. There, the emphasis will be on verbal creativity (poetry, humor), and is heavily supported by the data mining work on information networks.

We plan to build new contacts internationally to researchers in computational creativity as well as start collaboration on information browsing and access within HIIT (Prof. Jacucci).

Cooperation

Cooperation within Algodan

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

Cooperation within the universities

- Dr. Hannele Laivuori, Department of Medical Genetics: joint EU project on pre-eclampsia
- Dr. Outi Monni, Institute of Biomedicine: joint research in bioinformatics

Other national cooperation

Research collaboration with a number of companies in the Future Media and Data to Intelligence programmes of the ICT cluster of the Finnish Strategic Centres for Science, Technology and Innovation (Tivit Ltd).

International cooperation

- Professor Nada Lavrac, Jozef Stefan Institute, Slovenia: joint researech on data and text mining for bioinformatics. Joint publications.
- Professor Luc De Raedt, Katholieke Universiteit Leuven, Belgium: joint research on graph mining and statistical relational learning. Joint publications, researcher exchange.
- Professor Jiuyong Li, University of Southern Australia: joint research on dependency mining. Joint publications, research visits.

Societal, economical and technical impact

The Biomine search engine [2] is aiming at a societal impact in health care, where efficient access to medical information can facilitate both medical research and clinical health care.

Within computer science, the introduction of ProbLog, a probabilistic Prolog (in 2007, with the universities of Freiburg and Leuven, see also [10]) was milestone with an impact in the statistical relational learning community.

Finally, a whole different impact is aimed for in the field of computational creativity. We aim to publish computational poetry [5] and participate in public discussion on arts and creativity.

Selected publications

1. Jiuyong Li, Jiuxue Liu, Hannu Toivonen, and Jianming Yong. Effective pruning for the discovery of conditional functional dependencies. The Computer Journal. Accepted for publication.
2. Lauri Eronen and Hannu Toivonen. Biomine: Predicting links between biological entities using network models of heterogeneous database. BMC Bioinformatics, 2012. Accepted for publication.
3. Vid Podpecan, Nada Lavrac, Igor Mozetic, Petra Kralj Novak, Igor Trajkovski, Laura Langohr, Kimmo Kulovesi, Hannu Toivonen, Marko Petek, Helena Motaln, and Kristina Gruden. SegMine workflows for semantic microarray data analysis in Orange4WS. BMC Bioinformatics, 12(416), 2011.
4. Tuomas Tanner and Hannu Toivonen. Predicting and preventing student failure – using the k-nearest neighbour method to predict student performance in an online course environment. International Journal of Learning Technology, 5(4):356–377, 2010.

5. Jukka Toivanen, Hannu Toivonen, Alessandro Valitutti, and Oskar Gross. Corpus-based generation of content and form in poetry. In International Conference on Computational Creativity (ICCC), Dublin, Ireland, May-June 2012. To appear.
6. Laura Langohr and Hannu Toivonen. A model for mining relevant and non-redundant information. In ACM Symposium on Applied Computing (SAC), pages 451–456, Riva del Garda (Trento), Italy, March 2012.
7. Hannu Toivonen, Fang Zhou, Aleksi Hartikainen, and Atte Hinkka. Compression of weighted graphs. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), pages 965–973, San Diego, California, USA, August 2011.
8. Fang Zhou, Sebastien Mahler, and Hannu Toivonen. Network simplification with minimal loss of connectivity. In The 10th IEEE International Conference on Data Mining (ICDM), pages 659–668, Sydney, Australia, December 2010.
9. Petteri Hintsanen, Hannu Toivonen, and Petteri Sevon. Fast discovery of reliable subnetworks. In The 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 104–111, Odense, Denmark, 2010.
10. Luc De Raedt, Angelika Kimmig, Bernd Gutmann, Kristian Kersting, Vitor Santos Costa, and Hannu Toivonen. Probabilistic inductive querying using ProbLog. In S. Dzeroski, B. Goethals, and P. Panov, editors, Inductive Databases and Constraint-Based Data Mining, pages 229–262. Springer, 2010.
11. Alessandro Valitutti. Ambiguous Lexical Resources for Computational Humor Generation. In International Conference on Agents and Artificial Intelligence (ICAART), 2012.

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)
- Katja Astikainen, Doctoral student (-2011)
- Markus Heinonen, Doctoral student
- Panu Luosto, Doctoral student
- Hongyu Su, Doctoral student
- Yvonne Herrmann, MSc student (2011-2012)
- Huibin Shen, MSc student (2011-2012)
- Fitsum Tamene, MSc student (2012-)

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 biomarker discovery.

Research activities

L - learning from and mining structured and heterogeneous data(subgroup Kernel Machines, Pattern Analysis and Computational Biology)

In machine learning with structured outputs, we have developed new methods for multitask learning using max-margin conditional random field models. In particular, we have tackled the problem of multi-task drug bioactivity prediction (each attack is a cancer cell line) by inducing a graph between the tasks using auxiliary data and learning the labeling of the graph [2]. We have further developed an ensemble method where the final prediction is taken as a majority vote over a set of labeled graphs. We also show that it is possible to use random graphs as base classifiers, which opens the method to any multi-task or multi-label learning application [6].

Another problem of interest is efficient computation of graph kernels for molecular and chemical reaction data. In collaboration with prof. Veli Mäkinen's group, we developed a new method for path kernel computation making use of compressed indexes [1]. With that technology, we have computed path kernels for KEGG reaction database (ca. 10^4 entries ca. 10^8 kernel computations) and obtained state-of-the-art results in predicting enzyme classification from the chemical reaction description. Furthermore, we have developed new methods for predicting enzyme function under remote homology using so called reaction kernels. The method allows inter/extrapolation in the output space, and thus can postulate functions previously not characterized [5]. In mass spectrometric data analysis, we have recently developed kernel methods for de novo metabolite identification from tandem mass spectra [3]. Another example 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 kerneled [4].

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 [7]. Also, we have developed a new method that computes the atom-atom mappings in chemical reactions without solving the notorious subgraph isomorphism problem as a subtask [1].

F - Foundations of algorithmic data analysis (subgroup Learning Theory)

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.

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.
- Discovery of sparse patterns. In many applications, one is faced with the question how to extract statistical patterns involving a small set of variables from a large original data. Traditional feature selection methods often suffer from instability and inefficient computation. We focus on methods based on regularization with sparsity inducing norms, such as sparse PCA and CCA, which are potentially more stable and efficient for large data. Structural sparsity methods which allow prior information about the data to be introduced are of particular interest.
- Network prediction. 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).

Cooperation

Cooperation within Algodan

The group collaborates with the Succinct Data Structures group.

National cooperation

- University of Helsinki, Finland/Professor Liisa Holm. Collaboration in enzyme function prediction [5].
- University of Helsinki, CoE in Microbial Food Safety (Professors Airi Palva, Johanna Björkroth, Willem de Vos), collaboration in the analysis of biological networks in metagenomic microbe communities in food manufacturing processing.
- 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 up to now has lead to one published article and two manuscripts currently under review.

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.
- National Institute for Medical Research (NIMR), United Kingdom/Dr. Delmiro Fernandez-Reyes. Collaboration in biomarker discovery [4].
- ETH Zurich, Institue of Molecular Systems Biology, Dr. Nicola Zamboni. Collaboration in Mass Spectrometric data analysis.

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.

Selected publications

1. Heinonen, M., Lappalainen, S., Mielikäinen, T. J., & Rousu, J. (2011). Computing Atom Mappings for Biochemical Reactions without Subgraph Isomorphism. Journal of Computational Biology, 18(1), 43-58.
2. Su, H., Heinonen, M., & Rousu, J. (2010). Multilabel Classification of Drug-like Molecules via Max-Margin Conditional Random Fields. In 5th European Workshop on Probabilistic Graphical Models, pp. 265-272.
3. Heinonen, M., Välimäki, N., Mäkinen, V., & Rousu, J. (2012). Efficient Path Kernels for Reaction Function Prediction. In 3rd International Conference on Bioinformatics Models, Methods and Algorithms, Algarve, Portugal, February 2012, submitted.
4. Rousu, J., Agranoff, D., Shawe-Taylor, J., & Fernandez-Reyes, D. (2011). Sparse Canonical Correlation Analysis for Biomarker Discovery: A Case Study in Tuberculosis. In Machine Learning in Systems Biology: Proceedings of the Fifth International Workshop (pp. 73-77).

5. Astikainen, K., Holm, L., Pitkänen, E., Szedmak, S., & Rousu, J. (2011). Structured Output Prediction of Novel Enzyme Function with Reaction Kernels. In Biomedical Engineering Systems and Technologies. Springer Communications of Computer and Information Science 127(5), pp. 367-378
6. Rousu, J., & Su, H. (2011). Multi-Task Drug Bioactivity Classification with Graph Labeling Ensembles. In 6th International Conference on Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science 7036, pp. 157-167.
7. Pitkänen, E., Arvas, M., & Rousu, J. (2011). Minimum mutation algorithm for gapless metabolic network evolution. In 2nd International Conference Bioinformatics Models, Methods and Algorithms, Rome, January 2011.

Team Neuroinformatics

Neuroinformatics Group

Members

- Aapo Hyvärinen, Professor, Group leader
- Patrik Hoyer, Academy Research Fellow, Co-leader
- Michael Gutmann, Postdoctoral researcher
- Jun-ichiro Hirayama, Postdoctoral researcher (1/2010-6/2011)
- Hiroaki Sasaki, Postdoctoral researcher (10/2011-9/2012)
- Cristina Campi, Postdoctoral researcher (9/2010-8/2011)
- Hugo Eyeherabide, Postdoctoral researcher (10/2011-)
- Jukka-Pekka Kauppi, Postdoctoral researcher (6/2011-)
- Doris Entner, Doctoral student
- Antti Hyttinen, Doctoral student
- Jouni Puuronen, Doctoral student
- Miika Pihlaja, Masters/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 [1]. A matlab package was made available on the web as well.

Causal discovery. Our project on estimating causal relations from continuous-valued data was successful, and included 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 [2], a procedure to estimate the strength of causal effects in the presence of hidden variables [3], and identifiability results of linear cyclic causal models based on randomized experiments [4].

Neuroimaging data analysis. In our joint neuroimaging project with Riitta Hari, we investigated new variants of ICA for analysis of spontaneous activity (e.g. during rest) [5]. Also, we started working on "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. Combinations of ICA and inverse modelling in the context of MEG were another focus (publications under preparation for the last two topics).

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 [6] which shows a deep connection between supervised and unsupervised learning, and how it can be utilized to estimate such intractable models. The paper also shows how the method enables improvement of known results on modelling of natural image statistics. A general theoretical framework unifying these results with previous ones was proposed in [7].

Future plans

Two-person neuroscience. An important focus for the next couple of years will be the very deep project of two-person neuroscience, in collaboration with Prof. Riitta Hari of Aalto University. 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). No concrete results are available yet, but we are optimistic that when we figure out how to properly do all these three things, the results will have a great impact.

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 fromthe 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 (two months in 2011). One joint publication submitted.
- Frederick Eberhardt, Peter Spirtes, et al, Carnegie Mellon University, USA. Topic: theory of causal discovery. Several joint publications. Several visits (2-3 weeks each) by Patrik Hoyer in 2010-2012.
- 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, Japan. Topic: Natural image statistics. Joint publication.

Societal, economical and technical impact

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

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 S. M. Smith. Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. Under revision for J. of Machine Learning Research.
3. D. Entner, P. O. Hoyer, and P. Spirtes. Statistical test for consistent estimation of causal effects in linear non-Gaussian models. Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS-2012), La Palma, Canary Islands, 2012.
4. A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. Conditionally accepted to J. of Machine Learning Research.
5. 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, in press.
6. 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.
7. M. U. Gutmann and J. Hirayama. Bregman Divergence as General Framework to Estimate Unnormalized Statistical Models, Proc. Conf. on Uncertainty in Artificial Intelligence (UAI2011), 2011, Barcelona, Spain.

Publications

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.L. Eronen and H. Toivonen. Biomine: Predicting links between biological entities using network models of heterogeneous database. *BMC Bioinformatics*, 2012, to appear.
3. 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.
4. 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.
5. 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.
6. J. Li, J. Liu, H. Toivonen, and J. Yong. Effective pruning for the discovery of conditional functional dependencies. *The Computer Journal*, to appear.
7. 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. In *Proceedings of the Royal Society B*, 2012, published online in advance of the print journal.
8. P. Papapetrou, G. Benson and G. Kollios. Mining poly-regions in DNA. *International Journal of Data Mining and Bioinformatics (IJDMB)*, INDERSCIENCE, 2012, to appear.
9. M.A. Prada, J. Toivola, J. Kullaa and J. Hollmén. Three-way analysis of structural health monitoring data. *Neurocomputing*, 80, March 2012, pp. 119–128.
10. 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*, 2012, to appear.
11. L. Salmela. Average complexity of backward q-gram string matching algorithms. *Information Processing Letters*, 112(11), 2012, pp. 433–437.
12. O. Solin, E. Ukkonen and L. Haikala. Mining the UKIDSS GPS: star formation and embedded clusters. *Astronomy & Astrophysics*, 2012, to appear.
13. 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*, 2012, to appear.

Refereed conference articles and articles in edited books

14. 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, to appear.

15. 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, to appear.
16. M. Atkinson, J. Piskorski, H. Tanev, R. Yangarber and V. Zavarella. Techniques for Multilingual Security-related Event Extraction from Online News. In A. Przepiórkowski (editor), *Computational Linguistics—Applications,* Studies in Computational Intelligence, Springer-Verlag, 2012, to appear.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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)*, to appear.
23. 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, to appear.
24. 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, February, 2012.
25. 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), to appear.
26. 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, to appear.
27. S. Huttunen, A. Vihavainen, M. Du and R. Yangarber. Predicting Relevance of Event Extraction for the End User. In T. Poibeau et al. (editors), *Multi-source, Multilingual Information Extraction and Summarization, Theory and Applications of Natural Language Processing*, Springer-Verlag, 2012, to appear.
28. 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)*, to appear.

29. 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, to appear.
30. J. Kärkkäinen, S.J. Puglisi and D. Kempa. Slashing the Time for BWT Inversion. In *Proceedings of 2012 Data Compression Conference (DCC)*, IEEE Computer Society, pp. 99-108.
31. 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. 451-456.
32. L.A. Langohr and H. Toivonen. Retrieval of Relevant and Non-redundant Nodes. In *Proceedings of Workshop on Dynamic Network Analysis, in conjunction with Twelfth SIAM International Conference on Data Mining*, Anaheim, California, USA, April 2012.
33. 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, to appear.
34. 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, to appear.
35. A. Valitutti. Ambiguous Lexical Resources for Computational Humor Generation. In *Proceedings of International Conference on Agents and Artificial Intelligence (ICAART)*, 2012, to appear.
36. H. Wettig, K. Reshetnikov and R. Yangarber. Using Context and Phonetic Features in Models of Etymological Sound Change. In *Proceedings of EACL 2012: Workshop on Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, Avignon, France, 2012, to appear.

Technical reports and other publications

37. 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*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.
38. 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*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.
39. 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*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.
40. L.A. Langohr and H. Toivonen. Finding representative nodes in probabilistic graphs. In M.R. Berthold (editor), *Bisociative Knowledge Discovery*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.
41. A. Mozetic, N. Lavrac, V. Podpecan, P.K. Novak, H. Motaln, M. Petek, K. Gruden, H. Toivonen and K. Kulovesi. Bisociative knowledge discovery for microarray data analysis. In M.R. Berthold (editor), *Bisociative Knowledge Discovery*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.

42. 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, to appear.
43. H. Toivonen. Network Analysis: Overview. In M.R. Berthold (editor), *Bisociative Knowledge Discovery*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.
44. L. Eronen, H. Toivonen and P. Hintsanen. Biomine: A network-structured resource of biological entities for link prediction. In M.R. Berthold (editor), *Bisociative Knowledge Discovery*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.
45. 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*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.
46. F. Zhou, S.J. Mahler and H. Toivonen. Review of Network Abstraction Techniques. In M.R. Berthold (editor), *Bisociative Knowledge Discovery*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.
47. F. Zhou, S.J. Mahler and H. Toivonen. Simplification of Networks by Edge Pruning. In M.R. Berthold (editor), *Bisociative Knowledge Discovery*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7250, Springer-Verlag, 2012, to appear.

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

32. 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.
33. 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.
34. T. Ahonen. Kolmogorov Complexity in Lyrics. In *Proceedings of AdMIRe 2011*.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.

41. 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.
42. 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.
43. 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.
44. 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*.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.

53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.

67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.

80. 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.
81. 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.
82. 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.
83. 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

84. T. Ahonen, K. Lemström and S.M. Linkola. Compressing Quantized Tonal Centroid Vectors for Cover Song Identification. In *Proceedings of MIREX*, 2011.
85. 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.
86. 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.
87. 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.
88. 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.
89. V. Mäkinen. Algoritmitutkimuksen rooli bioinformatiikassa. *Tietojenkäsittelytiede*, 32, 2011, pp. 10–15.

2010

Articles in refereed scientific journals

1. E. Arkin, J. Mitchell and V. Polishchuk. Maximum Thick Paths in Static and Dynamic Environments. *Computational Geometry*, 43(3), 2010, pp. 279-294.
2. M. Arvas, N.S. Haiminen, B. Smit, J. Rautio, M. Vitikainen, M. Wiebe, D. Martinez, C. Chee, J. Kunkel, C. Sanchez, M.A. Nelson, N. Pakula, M. Saloheimo, M. Penttilä and T. Kivioja. Detecting novel genes with sparse arrays. *Gene*, 467 (1-2), 2010, pp. 41-51.
3. A. Bjorklund, T. Husfeldt, P. Kaski and M. Koivisto. Trimmed Moebius Inversion and Graphs of Bounded Degree. *Theory of Computing Systems*, 47(3), 2010, pp. 637-654.
4. A. Björklund, T. Husfeldt, P. Kaski and M. Koivisto. Evaluation of permanents in rings and semirings. *Information Processing Letters*, 110(20), 2010, pp. 867-870.

5. C.J. Colbourn, A.D. Forbes, M.J. Grannell, T.S. Griggs, P. Kaski, P.R.J. Östergård, D.A. Pike and O. Pottonen. Properties of the Steiner triple systems of order 19. *The Electronic Journal of Combinatorics*, 17(1), 2010, pp. R98.
6. J.T. Eronen, K. Puolamäki, L. Liu, K. Lintulaakso, J. Damuth, C. Janis and M. Fortelius. Precipitation and large herbivorous mammals, Part I: Estimates from present-day communities. *Evolutionary Ecology Research*, 12(2), 2010, pp. 217-233.
7. J.T. Eronen, K. Puolamäki, L. Liu, K. Lintulaakso, J. Damuth, C. Janis and M. Fortelius. Precipitation and large herbivorous mammals, Part II: Application to fossil data. *Evolutionary Ecology Research*, 12(2), 2010, pp. 235-248.
8. P. Floréen, P. Kaski, V. Polishchuk and J. Suomela. Almost Stable Matchings by Truncating the Gale–Shapley Algorithm. *Algorithmica*, 58(1), 2010, pp. 102-118.
9. N.S. Haiminen and H. Mannila. Evaluation of BIC and cross validation for model selection on sequence segmentations. *International Journal of Data Mining and Bioinformatics*, 4(6), 2010, pp. 675-700.
10. D. Hartley, N. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. Linge, A. Mawudeku, N. Collier, J. Brownstein, G. Thinus and N. Lightfoot. The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 2010/3, pp. 2-18.
11. H. Heikinheimo, J.T. Eronen, A. Sennikov, C. Preston, P. Uotila, H. Mannila and M. Fortelius. Converge in distribution patterns of Europe's plants and mammals is due to environmental forcing. *Journal of Biogeography, 2010*.
12. T. Honkela, A. Hyvärinen and J.J. Väyrynen. WordICA-emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16(3), 2010, pp. 277-308.
13. A. Hyvärinen, K. Zhang , S. Shimizu and P.O. Hoyer. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *Journal of Machine Learning Research*, 11, 2010, pp. 1709–1731.
14. A. Hyvärinen, P. Ramkumar, L. Parkkonen and R. Hari. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. *NeuroImage*, 49(1), 2010, pp. 257-271.
15. A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J.M. Vaquerizas, J. Yan, M.J. Sillanpää, A.W.M. Bonke, K. Palin, S. Talukder, T.R. Hughes, N.M. Luscombe, E. Ukkonen and J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, 20(6), 2010, pp. 861-873.
16. M. Korpela, H. Mäkinen, P. Nöjd, J. Hollmén and M. Sulkava. Automatic detection of onset and cessation of tree stem radius increase using dendrometer data. *Neurocomputing*, 73(10-12), June 2010, pp. 2039-2046.
17. U. Köster and A. Hyvärinen. A Two-Layer Model of Natural Stimuli Estimated with Score Matching. *Neural Computation*, 22(9), 2010, pp. 2308-2333.
18. K. Lemström, N. Mikkilä and V. Mäkinen. Filtering methods for content-based retrieval on indexed symbolic music databases. *Information retrieval*, 13(1), 2010, pp. 1-21.
19. M. Lukk, M. Kapushesky, J.T. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen and A. Brazma. A global map of human gene expression. *Nature Biotechnology*, 28(4), 2010, pp. 322-324.
20. 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), May 2010, pp. 1429-1450.

21. M. Michael, F. Nicolas and E. Ukkonen. On the complexity of finding gapped motifs. *Journal of Discrete Algorithms*, 8(2), 2010, pp. 131-142.
22. M. Ojala and G. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11, June 2010, pp. 1833-1863.
23. V. Mäkinen, G. Navarro, J. Sirén and N. Välimäki. Storage and Retrieval of Highly Repetitive Sequence Collections. *Journal of Computational Biology*, 17(3), 2010, pp. 281-308.
24. E. Pitkänen, J. Rousu and E. Ukkonen. Computational methods for metabolic reconstruction. *Current Opinion in Biotechnology*, 21(1), 2010, pp. 70-77.
25. J. Saarinen, E. Oikarinen, M. Fortelius and H. Mannila. The living and the fossilized: how well do unevenly distributed points capture the faunal information in a grid? *Evolutionary Ecology Research*, 12, 2010, pp. 363-376.
26. L. Salmela. Correction of sequencing errors in a mixed set of reads. *Bioinformatics*, 26(10), 2010, pp. 1284-1290.
27. T. Tanner and H. Toivonen. Predicting and preventing student failure - using the k-nearest neighbour method to predict student performance in an online course environment. *International Journal of Learning Technology*, 5(4), 2010, pp. 356-377.
28. E. Ukkonen. Geometric Point Pattern Matching in the Knuth-Morris-Pratt Way. *Journal of Universal Computer Science*, 16(14), 2010, pp. 1902-1911.
29. A. Usvasalo, E. Elonen, U.M. Saarinen-Pihkala, R. Räty, A. Harila-Saari, P. Koistinen, E.-R. Savolainen, S. Knuutila and J. Hollmén. Prognostic classification of patients with acute lymphoblastic leukemia by using copy number profiles identified from array-based comparative genomic hybridization data. *Leukemia Research*, 34(11), November 2010, pp. 1476-1482.
30. A. Usvasalo, S. Ninomiya, R. Räty, J. Hollmén, U.M. Saarinen-Pihkala, E.i Elonen, and S. Knuutila. Focal 9p instability in hematologic neoplasias revealed by comparative genomic hybridization and single nucleotide polymorphism microarray analyses. *Genes, Chromosomes, and Cancer*, 49(4), April 2010, pp. 309-318.
31. T. Vesala, S. Launiainen, P. Kolari, J. Pumpanen, S. Sevanto, P. Hari, E. Nikinmaa, P. Kaski, H. Mannila, E. Ukkonen, S.L. Piao and P. Ciais. Autumn temperature and carbon balance of a boreal Scots pine forest in Southern Finland. *Biogeosciences*, 7(1), 2010, pp. 163-176.
32. G. Wei, G. Badis, M.F. Berger, T. Kivioja, K. Palin, M. Enge, M. Bonke, A. Jolma, M. Varjosalo, A.R. Gehrke, J. Yan, S. Talukder, M. Turunen, M. Taipale, H.G. Stunnenberg, E. Ukkonen, T.R. Hughes, M.L. Bulyk and J. Taipale. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO Journal*, 29(13), 2010, pp. 2147-2160.
33. L. Yetukuri, J. Tikka, J. Hollmén and M. Orevic. Functional prediction of unidentified lipids using supervised classifiers. *Metabolomics*, 6(1), April 2010, pp. 18-26.
34. H. Yu and D.P. Bertsekas. Error Bounds for Approximations from Projected Linear Equations. *Mathematics of Operations Research*, 35(2), 2010, pp. 306-329.

Refereed conference articles and articles in edited books

35. P.R. Adhikari and J. Hollmén. Patterns from multiresolution 0-1 data. In *Proceedings of the ACM SIGKDD Workshop on Useful Patterns (UP '10)*, New York, NY, USA, 2010, ACM, pp. 8-16.
36. P. R. Adhikari and J. Hollmén. Preservation of statistically significant patterns in multiresolution 0-1 data. In T. Dijkstra, E. Tsivtsivadze, E. Marchiori and T. Heskes (editors), *Pattern Recognition in Bioinformatics*,2010, Lecture Notes in Computer Science, Springer-Verlag.

37. P. R. Adhikari and J. Hollmén. Mixture modelling of binary data. In *Statistical Mechanics of Learning and Inference*, 2010, Poster.

38. T. Ahonen. Combining Chroma Features for Cover Version Identification. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, Utrecht, The Netherlands, August, 2010, pp. 165-170.

39. T. Ahonen. Compressing Lists for Audio Classification. In *Proceedings of the 3rd International Workshop on Machine Learning and Music (MML 2010)*, Florence, Italy, October, 2010.

40. E. Arkin, A. Efrat, J. Mitchell, V. Polishchuk, S. Ramasubramanian, S. Sankararaman and J. Taheri. Data Transmission and Base-Station Placement for Optimizing Network Lifetime. In *DIALM-POMC '10 Proceedings of the 6th International Workshop on Foundations of Mobile Computing*, 2010, pp. 23-32.

41. D. Arroyuelo, F. Claude, S. Maneth, V. Mäkinen, G. Navarro, K. Nguyen, J.L.T. Siren and N. Välimäki. Fast In-Memory XPath Search using Compressed Indexes. In *Proceedings of the 26th IEEE International Conference on Data Engineering (ICDE 2010)*, Long Beach, USA, March, 2010, pp. 417-428.

42. K. Astikainen, E. Pitkänen, J. Rousu, L. Holm and S. Szedmak. Reaction Kernels: Structured Output Prediction Approaches for Novel Enzyme Function. In *Proceedings of the First International Conference on Bioinformatics*, 2010, pp. 48-55.

43. M. Atkinson, J. Piskorski, J. Belyaeva, S. Huttunen and R. Yangarber. Real-Time Text Mining in Multilingual News for the Creation of a Pre-frontier Intelligence Picture. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

44. M. Atkinson, J. Belyaeva, V. Zavarella, J. Piskorski, S. Huttunen, A. Vihavainen and R. Yangarber. News Mining for Border Security Intelligence. In *Proceedings of IEEE International Conference on Intelligence and Security Informatics (ISI),* 2010.

45. D.P. Bertsekas and H. Yu. Distributed Asynchronous Policy Iteration in Dynamic Programming. In *Proceedings of 2010 Allerton Conference on Communication, Control, and Computing*, 2010.

46. A. Björklund, T. Husfeldt, P. Kaski and M. Koivisto. Covering and packing in linear space. In *Proceedings of the 37th International Colloquium on Automata, Languages and Programming: Part I,* Bordeaux, France, July 6-10, 2010, Lecture Notes in Computer Science, Vol. 6198, Springer-Verlag, pp. 727-737.

47. L. De Raedt, A. Kimmig, B. Gutmann, K. Kersting, V. Santos Costa and H. Toivonen. Probabilistic Inductive Querying Using ProbLog. In S. Dzeroski, B. Goethals and P. Panov (editors), *Inductive Databases and Constraint-Based Data Mining*, 2010, Springer-Verlag, pp. 229-262.

48. B. Durian, H. Peltola, L. Salmela and J. Tarhio. Bit-parallel search algorithms for long patterns. In *Proceedings of the 9th International Symposium on Experimental Algorithms*, 2010, Lecture Notes in Computer Science, Vol. 6049, Springer-Verlag, pp. 129-140.

49. F. Eberhardt, P.O. Hoyer and R. Scheines. Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics,* 2010, pp. 185-192.

50. D. Entner and P.O. Hoyer. On causal discovery from time series data using FCI. In *Proceedings of the 5th European Workshop on Probabilistic Graphical Models*, 2010, pp. 121-128.

51. P. von Etter, S. Huttunen, A. Vihavainen, M. Vuorinen and R. Yangarber. Assessment of Utility in Web Mining for the Domain of Public Health. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, 2010.

52. U.M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *JMLR Workshop and Conference Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010),* 2010, pp. 297-304.
53. P. Hintsanen, H. Toivonen and P. Sevon. Fast Discovery of Reliable Subnetworks. In *Proceedings of 2010 International Conference on Advances in Social Network Analysis and Mining (ASONAM 2010)*, Odense, Denmark, 2010, pp. 104-111.
54. J. Hirayama, A. Hyvärinen and S. Ishii. Sparse and low-rank estimation of time-varying markov networks with alternating direction method of multipliers. In *Proceedings of the 17th International Conference on Neural Information Processing. Theory and Algorithms (ICONIP 2010)*, Part I, Sydney, Australia, November 22-25, 2010, pp. 371-379.
55. J. Hollmén, H. Mäkinen and P. Nöjd. Analyzing subjective expert opinions about standardization of tree-ring series. In *WorldDendro 2010 - Abstracts of the 8th International Conference on Dendrochronology*, Rovaniemi, Finland, June 2010, p. 110.
56. A. Hyttinen, F. Eberhardt and P.O. Hoyer. Causal discovery for linear cyclic models with latent variables. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM 2010)*, pp. 153-160.
57. A. Hyvärinen. Pairwise measures of causal direction in linear non-gaussian acyclic models. In *JMLR: Workshop and Conference Proceedings of 2nd Asian Conference on Machine Learning*, pp. 1-16.
58. D. Janzing, P.O. Hoyer and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.
59. M. Karvonen, M. Laitinen, K. Lemström and J. Vikman. Applying mathematical morphology for content-based music retrieval. In *Proceedings of International Symposium on Computer Music Modeling and Retrieval*, Malaga, Spain, 21-24 June, 2010.
60. M. Kasari, H. Toivonen and P. Hintsanen. Fast Discovery of Reliable k-terminal Subgraphs. In *Proceedings of Advances in Knowledge Discovery and Data Mining: The 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Part II, 2010, Lecture Notes in Artificial Intelligence, No. 6119, pp. 168-177.
61. M. Koivisto and P. Parviainen. A space-time tradeoff for permutation problems. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2010)*, 2010, SIAM, pp. 484–492.
62. W.M. Koolen, M.K. Warmuth and J. Kivinen. Hedging Structured Concepts. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT 2010)*, 2010, pp. 93-105.
63. M. Korpela, P. Nöjd, J.Hollmén, Harri Mäkinen, M. Sulkava and P. Hari. Daily temperature and daily photosynthetic production vs. Scots pine growth. In K. Mielikäinen, H. Mäkinen and M. Timonen (editors), *WorldDendro 2010 - Abstracts of the 8th International Conference on Dendrochronology,* Rovaniemi, Finland, June, 2010, pp. 261.
64. I. Kostitsyna and V. Polishchuk. Simple Wriggling is Hard unless you are a Fat Hippo. In *Fun with Algorithms*, 2010, Lecture Notes in Computer Science, Vol. 6099, pp. 272-283.
65. D.N. Krasnoshchekov, V., Polishchuk and A. Vihavainen. Shape approximation using k-order alpha-hulls. In *Proceedings of the 2010 Annual Symposium on Computational geometry (SoCG '10)*, 2010, pp. 109-110.
66. I. Krozel, J. Mitchell, V. Polishchuk and A. Pääkkö. Throughput/Complexity Tradeoffs for Routing Traffic in the Presence of Dynamic Weather. *In ICRAT 2010 Fourth International Conference on Research in Air Transportation*, 2010.

67. J. Kärkkäinen and S.J. Puglisi. Medium-Space Algorithms for Inverse BWT. In *Proceedings of 18th Annual European Symposium on Algorithms (ESA 2010)* , Part I, Liverpool, UK, September 6-8, 2010, Lecture Notes in Computer Science, Vol. 6346, Springer-Verlag, pp. 451-462.

68. G. Lejeune, A. Doucet, R. Yangarber and N. Lucas. Filtering news for epidemic surveillance: towards processing more languages with fewer resources. In *Proceedings of Fourth International Workshop on Cross Lingual Information Access at COLING 2010 (CLIA 2010)*, Beijing, China, 2010.

69. G. Lejeune, M. Hatmi, A. Doucet, S.M. Huttunen and N. Lucas. A proposal for a multilingual epidemic surveillance system. *In Revised Selected Papers of User Centric Media: First International Conference (UCMedia 2009)*, Venice, Italy, December 9-11, 2009, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, No. 17, Vol. 40, 2010, pp. 343-348.

70. K. Lemström. Towards More Robust Geometric Content-Based Music Retrieval. In *Proceedings of ISMIR 2010*, 2010, pp. 577-582.

71. K. Lemström. Transposition and time-scale invariant geometric music retrieval. In T. Elomaa, H. Mannila and P. Orponen (editors), *Algorithms and applications: Essays dedicated to Esko Ukkonen on the occasion of his 60th birthday*, Lecture Notes in Computer Science, Vol. 6060, Springer-Verlag, 2010.

72. J. Lijffijt, P. Papapetrou, and J. Hollmén. Tracking your steps on the track: Body sensor recordings of a controlled walking experiment. In *Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments (PETRA'2010),* June, 2010, ACM, article number 58.

73. J. Lijffijt, P. Papapetrou, J. Hollmén and V. Athitsos. Benchmarking dynamic time warping for music retrieval. In *Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments (PETRA'2010)*, June, 2010, ACM, article number 59.

74. J. Linge, R. Steinberger, F. Fuart, S. Bucci, J. Belyaeva, M. Gemo, D. Al-Khudhairy, R., Yangarber and E. van der Goot. MedISys: Medical Information System. In E Asimakopoulou (ed.), *Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed FrameworksIGI Global*,2010, pp. 131-142.

75. T. Lokki and K. Puolamäki. Canonical analysis of individual vocabulary profiling data. In *Proceedings of 2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, 2010, pp. 152-157.

76. P. Luosto, J. Kivinen and H. Mannila. Gaussian Clusters and Noise: An Approach Based on the Minimum Description Length Principle. In *Proceedings of 13th International Conference on Discovery Science (DS 2010)*, Canberra, Australia, October 6-8, 2010, Lecture Notes in Computer Science, Vol. 6332, pp. 251-265.

77. I. Mozetic, N. Lavrac, V. Podpecan, P.K. Novak, H. Motaln, M. Petek, K. Gruden, H. Toivonen and K. Kulovesi. Bisociative knowledge discovery for microarray data analysis. In *Proceedings of the International Conference on Computational Creativity (ICCC-X)*, January 7-9, 2010, Lisbon, Portugal, pp. 190-199.

78. V. Mäkinen, N. Välimäki, A. Laaksonen and R. Katainen. Unified view of backward backtracking in short read mapping. In T. Elomaa, H. Mannila and P. Orponen (editors), *Algorithms and applications: Essays dedicated to Esko Ukkonen on the occasion of his 60th birthday*, Lecture Notes in Computer Science, Vol. 6060, Springer-Verlag, 2010, pp. 182-195.

79. A. Norta, R. Yangarber and L. Carlson. Utility Evaluation of Tools for Collaborative Development and Maintenance of Ontologies. In *Proceedings of 14th IEEE International Enterprise Distributed Object Computing Conference Workshops (VORTE 2010/MOST 2010)*, 2010, pp. 207-214.

80. I. Nöllenburg, V. Polishchuk and M. Sysikaski. Dynamic One-Sided Boundary Labeling. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10)*, 2010, pp. 310-319.

81. M. Ojala. Assessing data mining results on matrices with randomization. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM '10)*, December, 2010, pp. 959-964.

82. M. Ojala, G. Garriga, A. Gionis and H. Mannila. Evaluating query result significance in databases via randomizations. In *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM '10)*, April, 2010, pp. 906-917.

83. P. Parviainen and M. Koivisto. Bayesian structure discovery in Bayesian networks with less space. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, JMLR: W&CP 9, Vol. 9, 2010, pp. 589–596.

84. M. Pihlaja, M.U. Gutmann and A.J. Hyvärinen. A Family of Computationally Efficient and Simple Estimators for Unnormalized Statistical Models. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, 2010.

85. V. Polishchuk and A. Vihavainen. Periodic Multi-Labeling of Public Transit Lines. In *Proceedings of 6th International Conference on GIScience (GIScience 2010)*, Zurich, Switzerland, September 14-17, 2010, Lecture Notes in Computer Science, Vol. 6292, Springer-Verlag, pp. 175-188.

86. M. A. Prada, J. Hollmén, J. Toivola and J. Kullaa. Three-way Analysis of Structural Health Monitoring Data. In S. Kaski, D.J. Miller, E. Oja and A. Honkela (editors), *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, August, 2010, IEEE, pp. 256-261.

87. K. Puolamäki, A. Bertone, R. Theron, O. Huisman, J. Johansson, S. Miksch, P. Papapetrou and S. Rinzivillo. Data Mining. In D. Keim, J. Kohlhammer, G. Ellis and F. Mansmann (editors), *Mastering the Information Age Solving Problems with Visual Analytics*, Chapter 4, Eurographics Association, 2010, pp. 39-56.

88. K. Puolamäki, P. Papapetrou and J. Lijffijt. Visually controllable data mining methods. In *Proceedings of IEEE International Conference on Data Mining Workshops 2010*, 2010, pp. 409-417.

89. P. Ramkumar, A. Hyvärinen, L. Parkkonen and R. Hari. Characterization of spontaneous neuromagnetic brain rhythms using independent component analysis of short-time Fourier transforms. In *IFMBE Proceedings of 17th International Conference on Biomagnetism Advances in Biomagnetism (Biomag 2010)*, Dubrovnik, Croatia, March 28–April 1, 2010.

90. L. Salmela and J. Tarhio. Approximate string matching with reduced alphabet. In T. Elomaa, H. Mannila and P. Orponen (editors), *Algorithms and applications: Essays dedicated to Esko Ukkonen on the occasion of his 60th birthday*, Lecture Notes in Computer Science, Vol. 6060, Springer-Verlag, 2010.

91. J. Sirén. Sampled Longest Common Prefix Array. In *Proceedings of 21st Annual Symposium on Combinatorial Pattern Matching (CPM 2010)*, New York, USA, June 2010, Lecture Notes in Computer Science, Vol. 6129, Springer-Verlag, pp. 227-237.

92. K. Sirvio and J. Hollmén. Multi-year network level road maintenance programming by genetic algorithms and variable neighbourhood search. In *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*, September, 2010, pp. 581-586.

93. Y. Sogawa, S. Shimizu, A. Hyvärinen, T. Washio, T. Shimamura and S. Imoto. Discovery of exogenous variables in data with more variables than observations. In *Proceedings of 20th International Conference on Artificial Neural Networks (ICANN 2010),* Part I, Thessaloniki, Greece, September 15-18, 2010, Lecture Notes in Computer Science, Vol. 6352, Springer-Verlag, pp. 67-76.

94. H. Su, M. Heinonen and J. Rousu. Multilabel Classification of Drug-like Molecules via Max-Margin Conditional Random Fields. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM-2010)*, 2010, pp. 265-272.

95. H. Su, M. Heinonen and J. Rousu. Structured Output Prediction of Anti-Cancer Drug Activity. In *Pattern Recognition in Bioinformatics*, 2010, Lecture Notes in Computer Science, Vol. 6282, Springer-Verlag, pp. 38-49.

96. M. Timonen, P. Silvonen and M. Kasari. Modelling a Query Space Using Associations. In *Proceedings of the 20th European-Japanese Conference on Information Modelling and Knowledge Bases*.

97. J. Toivola, M.A. Prada and J. Hollmén. Novelty detection in projected spaces for structural health monitoring. In *Proceedings of the 9th International Symposium on Advances in Intelligent Data Analysis (IDA 2010)*, Tuscon, Arizona, May 2010, Lecture Notes in Computer Science, Vol. 6065, Springer-Verlag, pp. 208-219.

98. H. Toivonen, S. Mahler and F. Zhou. A Framework for Path-Oriented Network Simplification. In *Proceedings of 9th International Symposium on Advances in Intelligent Data Analysis (IDA 2010)*, Tuscon, Arizona, May 2010, Lecture Notes in Computer Science, Vol. 6065, Springer-Verlag, pp. 220-231.

99. A. Ukkonen. The support vector tree. In *Algorithms and Applications*, 2010, pages 244-259.

100. A. Ukkonen and M. Arias. Example-dependent basis vector selection for kernel-based classifiers. In *Proceedings of ECML/PKDD*, 2010, pp. 338-353.

101. N. Vuokko. Consecutive ones property and spectral ordering. In *Proceedings of the 10th SIAM International Conference on Data Mining (SDM '10)*, 2010, pp. 350-360.

102. N. Vuokko and P. Kaski. Testing the significance of patterns in data with cluster structure. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM2010)*, Sydney, Australia, December 14—17, 2010, pp. 1097-1102.

103. N. Välimäki, S. Ladra and V. Mäkinen. Approximate All-Pairs Suffix/Prefix Overlaps. In *Proceedings of 21st Annual Symposium on Combinatorial Pattern Matching (CPM 2010)*, New York, USA, June 2010, Lecture Notes in Computer Science, Vol. 6129, Springer-Verlag, pp. 76-87.

104. J. Wettig, S. Hiltunen and R. Yangarber. Hidden Markov Models for Induction of Morphological Structure of Natural Language. In *Proceedings of Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2010)*, Tampere, Finland, August 2010.

105. H. Yu. Convergence of Least Squares Temporal Difference Methods under General Conditions. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

106. V. Zarzoso and A.J. Hyvärinen. Iterative algorithms. In *Handbook on Independent Component Analysis and Blind Source Separation*, Academic Press, 2010.

107. K. Zhang and A. Hyvärinen. Source separation and higher-order causal analysis of MEG and EEG. In *Proceedings of the Twenty-Sixth Conference Uncertainty in Artificial Intelligence (UAI 2010)*, 2010, pp. 709-716.

108. K. Zhang and A. Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *JMLR Workshop and Conference Proceedings: Causality: Objectives and Assessment (NIPS 2008*), Volume 6, 2010, pp. 157–164.

109. F. Zhou, S.J. Mahler and H. Toivonen. Network Simplification with Minimal Loss of Connectivity. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010)*, Sydney, Australia, December, 2010, pp. 659-668.

Technical reports and other publications

110. M. Arita, M. Heinonen and J. Rousu, editors. *Mass Spectrometry Informatics in Systems Biology: Abstracts of the Workshop.* Series of Publications C, University of Helsinki, Department of Computer Science, 2010.
111. D. Bertsekas and H. Yu. *Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming (Revised).* Technical Report, C-2010-10, Series of Publications C, University of Helsinki, Department of Computer Science, 2010.
112. T. Elomaa, H. Mannila and P. Orponen, editors. *Algorithms and Applications: Essays Dedicated to Esko Ukkonen on the Occasion of His 60th Birthday,* Lecture Notes in Computer Science, Vol. 6060, Springer-Verlag, 2010.
113. A. Hyvärinen. Statistical models of natural images and cortical visual representation. *Topics in Cognitive Science*, 2(2), 2010, pp. 251-264.
114. J. Lijffijt, P. Papapetrou, N. Vuokko and K. Puolamäki. *The smallest set of constraints that explains the data: a randomization approach*. Technical Report TKK-ICS-R31, Aalto University School of Science and Technology, Department of Information and Computer Science, Espoo, Finland, May 2010.
115. L. Salmela. Merkkijonoalgoritmeja monen hahmon hakuun. *Tietojenkäsittelytiede*, 31, 2010, pp. 70–83.
116. H. Toivonen. Frequent Pattern. In C. Sammut and G.I. Webb (editors), *Encyclopedia of Machine Learning*, 1st edition, Springer-Verlag, 2010.
117. H. Toivonen. Apriori Algorithm. In C. Sammut and G.I. Webb (editors), *Encyclopedia of Machine Learning*, 1st edition, Springer-Verlag, 2010.
118. H. Toivonen. Association Rule. In C. Sammut and G.I. Webb (editors), *Encyclopedia of Machine Learning*, 1st edition, Springer-Verlag, 2010.
119. H. Toivonen. Basket Analysis. In C. Sammut and G.I. Webb (editors), *Encyclopedia of Machine Learning*, 1st edition, Springer-Verlag, 2010.
120. H. Toivonen. Frequent Itemset. In C. Sammut and G.I. Webb (editors), *Encyclopedia of Machine Learning*, 1st edition, Springer-Verlag, 2010.
121. A. Ukkonen. Approximate top-k retrieval from hidden relations. *CoRR*, abs/1008.5057, 2010.
122. E. Ukkonen. Tila ja tulevaisuus: pysähtyneisyydestä uuteen vauhtiin. *Tietojenkäsittelytiede*, 30, 2010, pp. 4-6.
123. H. Yu. *Convergence of Least Squares Temporal Difference Methods under General Conditions.* Report, C-2010-1, Series of Publications C, University of Helsinki, Department of Computer Science, 2010.
124. H. Yu. *Least Squares Temporal Difference Methods: An Analysis under General Conditions.* Technical Report, C-2010-39, Series of Publications C, University of Helsinki, Department of Computer Science, 2010.

PhD degrees

Algodan also funds Finnish graduate students working in Algodan research groups. Below is a list of PhD degrees obtained by Algodan researchers.

2012

1. Hanhijärvi, Sami. Multiple hypothesis testing in data mining. *Aalto University.*
2. Parviainen, Pekka. Algorithms for Exact Structure Discovery in Bayesian Networks. *University of Helsinki.*
3. Vuokko, Niko. Testing the Significance of Patterns in Complex Null Hypotheses*. Aalto University.*
4. Wessman, Jaana. Mixture Model Clustering in the Analysis of Complex Diseases. *University of Helsinki.*

2011

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

2010

8. Heikinheimo, Hannes. Extending data mining techniques for frequent pattern discovery: trees, lowentropy sets, and crossmining. *Aalto University.*
9. Hämäläinen, Wilhelmiina. Efficient search for statistically significant dependency rules in binary data. *University of Helsinki.*
10. Kollin, Jussi. Computational Methods for Detecting Large-Scale Chromosome Rearrangements in SNP Data. *University of Helsinki.*
11. Lukk, Margus. Construction of a global map of human gene expression – the process, tools and analysis. *University of Helsinki.*
12. Pitkänen, Esa. Computational Methods for Reconstruction and Analysis of Genome-Scale Metabolic Networks. *University of Helsinki.*

Funding

Funding agency	2008	2009	2010	2011
Academy of Finland	1 548 00	1 319 000	1 518 000	1 604 000
Tekes	46 000	223 000	215 000	5 000
EU	362 000	340 000	59 000	69 000
Other	79 000	212 000	241 000	13 000
Own funding	1 055 000	1 324 000	1 452 000	1 103 000
Ministry	2 000	66 000		
TOTAL	3 092 007	3 484 000	3 485 000	2 794 000