Program
of the Sixth International Workshop on

Machine Learning in Systems Biology

(MLSB 2012)

in Basel, Switzerland

on September 8th & 9th, 2012

Sponsor: PASCAL2 Network of Excellence

Preface

Biology is rapidly turning into an information science, thanks to enormous advances in the ability to observe the molecular properties of cells, organs and individuals. Today's data allows us to model molecular systems at an unprecedented level of detail and to start to understand the underlying biological mechanisms. This wealth of data had a significant impact on the field of Systems Biology: There is an ever increasing demand for Machine Learning methods that identify statistical dependencies and patterns in large-scale datasets and that model complex biological phenomena on the molecular as well as systems level.

Our workshop is the sixth installment of the meeting on Machine Learning in Systems Biology (MLSB). We are honored and are excited to provide again the opportunity to exchange ideas and results at the interface of Machine Learning and Systems Biology. MLSB 2012 will feature four invited talks by leading experts in the field: Yves Moreau, Pascal Falter-Braun, Uwe Ohler, and Ben Lehner. For the first time, MLSB is accompanied by a virtual issue in OUP Bioinformatics. We received a tremendous response of 36 high-quality full paper submissions. In addition we received 21 submissions of extended abstracts. In a very competitive reviewing and selection process, we have selected the best pieces to be presented orally at the meeting. In addition we provide opportunities for poster presentations. We are looking forward to two days of interesting talks and posters.

We would like to thank Max Zwießele and Bijan Azmoun for their help with this program booklet, and the program committee members and reviewers for their careful and rapid reviewing for MLSB and for the Bioinformatics virtual issue.

These are exciting times for researchers working at the interface of Biology and Computer Science. We hope that MLSB 2012 in Basel may be a constructive and inspiring opportunity for these two fields. We encourage the open exchange of ideas as well as data and hope to create new interdisciplinary collaborations. We wish you an exciting and scientifically stimulating meeting in Basel!

Karsten Borgwardt, Max Planck Institutes and University of Tübingen, Germany
Gunnar Rätsch, Memorial Sloan-Kettering Cancer Center, New York, USA

Boston and New York, July 2012

MLSB 2012 Program

Basel, Switzerland, September 8th and 9th, 2012

September 8, 2012, Congress Center Basel, Workshop WS1

Registration

Session 1: 8:50am-10:30am

- Introduction
- **Invited Talk**: **Uwe Ohler**, Duke University, *Deciphering transcription regulation: from individual sites to cell type specific expression*
- **Jianlong Qi**: *Context-specific transcriptional regulatory network inference from global gene expression maps using double two-way t-tests*

Coffee Break: 10:30am-11:00am

Session 2: 11:00am-12:30pm

- **Luna De Ferrari**: *Active and guided learning of enzyme function*
- **Joeri Ruyssinck**: *Inferring gene regulatory networks using ensembles of feature selection techniques*
- **Thais G. do Rego**: *Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models*

Lunch: 12:30pm-1:30pm

Session 3: 1:30pm-3:00pm

- **Invited Talk**: **Yves Moreau**, KU Leuven, *Kernel methods for genomic data fusion*
- **Bettina Knapp**: *Efficient, data-based network inference using a linear programming approach*

Poster Session 3:00pm-4:30pm

Session 4: 4:30pm-6:00pm

- **Celine Brouard**: *Learning a Markov logic network for supervised gene regulation inference: application to the ID2 regulatory network in human keratinocytes*
- **Federica Eduati**: *Integrating literature-constrained and data-driven inference of signalling networks*
- **Daniela Stojanova**: *Using PPI Networks in hierarchical multi-label classification trees for gene function prediction*

September 9, 2012, Congress Center Basel, Workshop WS1

Session 5: 9:00am-10:30am

- **Invited Talk: Pascal Falter-Braun,** TU Munich, *Signatures of evolution and systems organization from an Arabidopsis interactome network map*
- **Paurush Praveen**: *Boosting statistical network inference by incorporating prior knowledge from multiple sources*

Coffee Break: 10:30am-11:00am

Session 6: 11:00am-12:30pm

- **Mehmet Gönen**: *Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization*
- **Elisabeth Georgii**: *Targeted retrieval of gene expression measurements using regulatory models*
- **Markus Heinonen**: *Metabolite identification and molecular fingerprint prediction via machine learning*

Lunch: 12:30pm-1:30pm

Session 7: 1:30pm-3:00pm

- **Invited Talk**: **Ben Lehner**, Centre for Genomic Regulation Barcelona, *The genetics of individuals: why would a mutation kill me, but not you?*
- **Chris Oates**: *Network inference using steady-state data and Goldbeter-Koshland kinetics*

Coffee Break 3:00pm-3:30pm

Session 8: 3:30pm-4:30pm

- **Thomas Sakoparnig**: *Efficient sampling for Bayesian inference of conjunctive Bayesian networks*
- **Felix Sanchez-Garcia**: *Helios: discovering driver oncogenes*

Closing Remarks: 4:30pm

Invited Talks

Deciphering transcription regulation: from individual sites to cell type specific expression

Uwe Ohler, Duke University & Berlin Institute for Medical Systems Biology/Max Delbrück Center & Humboldt Universität Berlin

Understanding how transcription regulation is encoded in the genomes of complex multicellular organisms has been a big challenge, not least due to the large non-coding space where relevant interactions might occur. High throughput technologies now allow it to map putative regulatory regions via their chromatin structure, and have made rapid progress in identifying in vivo binding of transcription factors to DNA at high resolution. Large collections of relevant data have been made available by individual groups as well as large consortia such as mod/ENCODE. I will discuss some of our recent and ongoing efforts that make use of such datasets to define successful computational models for individual sites as well as for cell-type specific expression.

Kernel methods for genomic data fusion

Yves Moreau, KU Leuven, Belgium

Despite significant advances in omics techniques, the identification of genes causing rare genetic diseases and the understanding of the molecular networks underlying those disorders remains difficult. Gene prioritization attempts to integrate multiple, heterogeneous data sources to identify candidate genes most likely to be associated with or causative for a disorder. Such strategies are useful both to support clinical genetic diagnosis and to speed up biological discovery. Genomic data fusion algorithms are rapidly maturing statistical and machine learning techniques have emerged that integrate complex, heterogeneous information (such as sequence similarity, interaction networks, expression data, annotation, or biomedical literature) towards prioritization, clustering, or prediction. In this talk, we will focus in particular on kernel methods and will propose several strategies for prioritization and clustering in particular. We also go beyond learning methods as such by addressing how such strategies can be embedded into the daily practice of geneticists, mostly through collaborative knowledge bases that integrate tightly with prioritization and network analysis methods.

Signatures of evolution and systems organization from an *Arabidopsis* interactome network map

Pascal Falter-Braun, TU München, Germany

Elucidating mechanisms of life requires analysis of whole systems and understanding the complex interplay of the individual components. Proteins control and mediate the majority of biological activities and interactions among proteins play a decisive role in the dynamic modulation of cellular behavior. Protein-protein interactions are essential constituents of all cells and interactome analysis is an important component in the quest for a systems level understanding of life.

We explore interactome networks for yeast, human and plant at ever increasing completeness and quality using both experimental and computational mapping and analysis tools. Based on benchmarking and standardized reference sets we have developed experimental approaches and mathematical models for the quantitative evaluation of the completeness and quality of interactome maps. These models enable a critical assessment of current maps and guide development of a roadmap towards completion.

Recently mapping of the first binary interactome network for the reference plant *Arabidopsis thaliana* was completed. Using tools of graph theory we identify biologically relevant network communities from which a picture of the overall interactome network organization starts to emerge. Combination of interaction and comparative genomics data yielded insights into network evolution, and biological inspection resulted in many hypotheses for unknown proteins and revealed unexpected connectivity between previously studied components of phytohormone signaling pathways.

Using the network we explored how bacterial and fungal pathogens perturb their host's network. Pathogen effectors from evolutionary distant pathogens were found to converge on network hubs, which appear "guarded" by resistance proteins, and which we show to be functionally important for the host's immune responses. Genetically, we were able to validate >90% of the Arabidopsis proteins targeted by both pathogens. Together, we show how high-quality protein interactome network maps provide us with tools for elucidating fundamental laws underlying biological systems.

The genetics of individuals: why would a mutation kill me, but not you?

Ben Lehner, Centre for Genomic Regulation, Barcelona, Spain

To what extent is it possible to predict the phenotypic differences among individuals from their completely sequenced genomes? We use model organisms (yeast, worms) to understand when you can, and why you cannot, predict the biology of an individual from their genome sequence.

Full oral presentations

Context-specific transcriptional regulatory network inference from global gene expression maps using double two-way t-tests

Jianlong Qi[1] and Tom Michoel[2]
[1] Center for Comparative Genomics and Bioinformatics, United States; [2] Freiburg Institute for Advanced Studies (FRIAS), Germany

Motivation: Transcriptional regulatory network inference methods have been studied for years. Most of them relie on complex mathematical and algorithmic concepts, making them hard to adapt, re- implement or integrate with other methods. To address this problem, we introduce a novel method based on a minimal statistical model for observing transcriptional regulatory interactions in noisy expression data, which is conceptually simple, easy to implement and integrate in any statistical software environment, and equally well performing as existing methods.
Results: We developed a method to infer regulatory interactions based on a model where transcription factors (TFs) and their targets are both differentially expressed in a gene-specific, critical sample contrast, as measured by repeated two-way t-tests. Benchmarking on standard E. coli and yeast reference datasets showed that this method performs equally well as the best existing methods. Ana- lysis of the predicted interactions suggested that it works best to infer context-specific TF-target interactions which only co-express locally. We confirmed this hypothesis on a dataset of more than 1,000 normal human tissue samples, where we found that our method predicts highly tissue-specific and functionally relevant interactions, whereas a global co-expression method only associates general TFs to non-specific biological processes.
Availability: A software tool called TwixTrix is available from http://omics.frias.uni-freiburg.de/software. Supplementary Material is available from http://omics.frias.uni-freiburg.de/supplementary-data.
Contact: tom.michoel@roslin.ed.ac.uk

Active and Guided Learning of Enzyme Function

Luna De Ferrari[1], Stuart Aitken[2], and John Mitchell[3].
[1] University of St Andrews, United Kingdom; [2] University of Edinburgh, United Kingdom; [3] University of St Andrews, United Kingdom.

Manual annotation cannot keep up with enzyme sequence discovery. In this work, we modelled the use of active and guided learning to support enzyme function curation. We evaluated, on 5,750 E. coli proteins, nine strategies to sort instances for curation. We found that selecting sets of InterPro features in order of frequency of occurrence can cut the curation effort by almost two thirds, while maintaining very high accuracy and recall. The method can be applied to real-life datasets of millions of proteins thanks to its limited computational requirements, parallelisation, good coverage of rare classes and flexibility in selecting instances for annotation.

Inferring gene regulatory networks using ensembles of feature selection techniques

Joeri Ruyssinck[1], Vân anh Huynh-thu[2], Pierre Geurts[2], Tom Dhaene[*], Piet Demeester[1], Yvan Saeys[3]
[1] Department of Information Technology, Ghent University – IBBT, Gent, Belgium, [2] Department of Electrical Engineering and Computer Science, Systems and Modeling, University of Liège, Liège, Belgium, [3] Department of Plant Systems Biology VIB, Ghent University, Gent, Belgium

Motivation: One of the long-standing open challenges in computational systems biology is the inference of gene regulatory networks from expression data. Recently, two community-wide efforts have been established to benchmark network inference techniques (DREAM4 and DREAM5), where it was shown that a feature selection method us- ing variable importance scores obtained from

tree-based ensemble methods (GENIE3) achieved top performance. Despite the success of this algorithm, little research has been carried out to understand why this approach works so well, and if equally good or better results could be obtained using other types of feature selection techniques.
Results: In this work, we present a large scale analysis of feature selection approaches to the network inference problem. We show that, using the recent concept of ensemble feature selection techniques, equally good or better results than GENIE3 can be achieved, demonstrating that the ensemble setting is a necessary requirement for feature selection techniques to achieve good performance on the network inference task. Furthermore, we show that by combining several ensemble feature selection techniques the performance can be made more robust and slightly improved. This analysis opens up new avenues for the development of novel types of ensemble based feature selection techniques in this setting.
Availability: R source code of all methods can be downloaded at
http://studwww.ugent.be/~jruyssun/FS/
Contact: joeri.ruyssinck@ugent.be

Inferring Epigenetic and Transcriptional Regulation during Blood Cell Development with a Mixture of Sparse Linear Models

Thais G. do Rego[1], Helge G. Roider[2], Francisco A. T. de Carvalho[1], and Ivan G. Costa[1]
[1] Center of Informatics, Federal University of Pernambuco, Recife, Brazil, [2] Merck Serono, Germany

Motivation: Blood cell development is thought to be controlled by a circuit of transcription factors and chromatin modifications that deter- mine the cell fate via activating cell type specific expression programs. To shed light on the interplay between histone marks and transcrip- tion factors during blood cell development, we model gene expression from regulatory signals by means of combinations of sparse linear regression models.
Results: The mixture of sparse linear regression models was able to improve the gene expression prediction in relation to the use of a single linear model. Moreover, it performed an efficient selection of regulatory signals even when analyzing all transcription factors with known motifs (> 600). The method identified interesting roles for histone modifications and a selection of transcription factors related to blood development and chromatin remodelling.
Availability: The method and data sets are available from http://www.cin.ufpe.br/˜igcf/SparseMix.
Contact: igcf@cin.ufpe.br

Efficient, Data-Based Network Inference using a Linear Programming Approach

Bettina Knapp[1], Johanna Mazur[1,2], and Johanna Kaderali[1,2]
[1] Heidelberg University, ViroQuant Research Group Modeling, BioQuant BQ26, Im Neuenheimer Feld 267, Heidelberg, Germany, [2] Dresden University of Technology, Medical Faculty Carl Gustav Carus, Institute for Medical Informatics and Biometry, Fetscherstrasse 74, Dresden, Germany

Motivation: In the recent years, technical developments enabled the facilitated measurements of biological high-throughput data. This results in a qualitative and a quantitative improvement of the gene- rated data and offers the potential to understand complex biological systems in more detail. Perturbation experiments, for example using RNA interference, are an easy and fast way to screen genes in a high-content, high-throughput manner and thereby, to elucidate their gene function. The inference of signal transduction networks from this data, however, is a challenging task. One of the problems is the exponentially increasing number of possible network topologies with an increasing number of nodes. Here, we formulate the problem of net- work inference as a linear optimization program which can be solved efficiently even for large-scale problems.
Results: Based on simulated data for networks of different sizes we show that our method outperforms a recently published approach, especially when applied to large-scale problems. Using our approach, we achieve increased sensitivity and specificity values and a signi- ficant reduction in computation time in comparison to the other approach. Furthermore, we show that our method can deal with noisy and missing data and that prior knowledge can be easily integrated and thus, improves results. Finally, we use a real data set studying ErbB signaling to reconstruct the

underlying network topology. Based on the gene interactions as given in the STRING database we achieve an accuracy much better than random guessing. We were able to reconstruct several already known interactions, as well as identify potential new ones.
Availability: The R source code of the method can be downloaded from http://tu-dresden.de/med/lpmodel
Contact: bettina.knapp@tu-dresden.de

Learning a Markov Logic Network for supervised gene regulation inference: application to the ID2 regulatory network in human keratinocytes

Celine Brouard[1], Julie Dubois[1,2], Christel Vrain[2], David Castel[3], Marie-anne Debily[3], Florence d'Alché-Buc[1,4]
[1] d'Orléans University LIFO, France, [2] CEA, France, [3] IBISC University of Evry, France

Motivation: Gene regulatory network inference remains a challenging problem in systems biology despite numerous approaches. When substantial knowledge on a gene regulatory network is already available, supervised network inference also is appropriate. Such a method builds a binary classifier able to assign a class (Regulation/No regulation) to an ordered pair of genes. Once learnt, the classifier can be used to predict new regulations. In this work, we explore the framework of Markov Logic Network (MLN) recently introduced by Richardson & Domingos (2004, 2006). A MLN is a random Markov network that codes for a set of weighted formula. It therefore combines features of probabilistic graphical models with the expressivity of 1st order logic rules.
Results: Starting from a known gene regulatory network involved in the switch proliferation differentiation of keratinocytes cells, a set of experimental transcriptomic data, and description of genes in terms of GO terms encoded into first order logic, we learn a Markov Logic network, e.g. a set of weighted rules that conclude on the predicate ”regulates”. As a side contribution, we define a list of basic tests for performance assessment, valid for any binary classifier. A first test consists of measuring the average performance on balanced edge prediction problem; a 2nd one deals with the ability of the classifier, once enhanced by asymmetric bagging, to update a given network; finally a 3rd test measures the ability of the method to predict new interactions with new genes.
Conclusion: The numerical studies show that MLNs achieve very good prediction while opening the door to some interpretability of the decisions. Additionally to the ability to suggest new regulation, such an approach allows to cross-validate experimental data with existing knowledge.
Availability: The code will be available on demand.
Contact: celine.brouard@ibisc.univ-evry.fr, florence.dalche@ibisc.univevry.fr

Integrating literature-constrained and data-driven inference of signalling networks

Federica Eduati[1], Javier De Las Rivas[2], Barbara DiCamillo[1], Gianna Toffolo[1] and Julio Saez-Rodriguez[3]
[1] University of Padova, Italy; [2] Bioinformatics and Functional Genomics Research Group Cancer Research Center (CiC-IBMCC, CSIC/USAL), Spain; [3] European Bioinformatics Institute, UK

Motivation: Recent developments in experimental methods allow generating increasingly larger signal transduction datasets. Two main approaches can be taken to derive from these data a mathematical model: to train a network (obtained e.g. from literature) to the data, or to infer the network from the data alone. Purely data-driven methods scale up poorly and have limited interpretability, while literature-constrained methods cannot deal with incomplete networks.
Results: We present an efficient approach, implemented in the R package CNORfeeder, to integrate literature-constrained and data- driven methods to infer signalling networks from perturbation experiments. Our method extends a given network with links derived from the data via various inference methods, and uses information on physical interactions of proteins to guide and validate the integration of links. We apply CNORfeeder to a network of growth and inflammatory signaling, obtaining a model with superior data fit in the human liver cancer HepG2 and proposes potential missing pathways.

Availability: CNORfeeder is in the process of being submitted to Bioconductor and in the meantime available at www.ebi.ac.uk/~cokelaer/cnofeeder/.
Contact: saezrodriguez@ebi.ac.uk

Using PPI Networks in Hierarchical Multi-label Classification Trees for Gene Function Prediction

Daniela Stojanova[1], Michelangelo Ceci[2], Donato Malerba[2] and Saso Dzeroski[1]
[1] Jozef Stefan Institute, Slovenia; [2] University of Bari, Italy

Motivation: Catalogs, such as Gene Ontology (GO) and MIPS-FUN, assume that functional classes are organized hierarchically (general functions include more specific functions). This has recently motivated the development of several machine learning algorithms under the assumption that instances may belong to multiple hierarchy organized classes. Besides relationships among classes, it is also possible to identify relationships among examples. Although such relationships have been identified and extensively studied in the in the area of protein-to-protein interaction (PPI) networks, they have not received much attention in hierarchical protein function prediction. The use of such relationships between genes introduces autocorrelation and violates the assumption that instances are independently and identically distributed, which underlines most machine learning algorithms. While this consideration introduces additional complexity to the learning process, we expect it would also carry substantial benefits.
Results: This article demonstrates the benefits (in terms of predictive accuracy) of considering autocorrelation in multi-class gene function prediction. We develop a tree-based algorithm for considering network autocorrelation in the setting of Hierarchical Multi-label Classification (HMC). The empirical evaluation of the proposed algorithm, called NHMC, on 24 yeast datasets using MIPS-FUN and GO annotations and exploiting three different PPI networks, clearly shows that taking autocorrelation into account improves performance.
Conclusions: Our results suggest that explicitly taking network autocorrelation into account increases the predictive capability of the models, especially when the underlying PPI network is dense. Furthermore, NHMC can be used as a tool to assess network data and the information it provides with respect to the gene function.

Boosting Statistical Network Inference by Incorporating Prior Knowledge from Multiple Sources

Paurush Praveen[1] and Holger Fröhlich[1]
[1] University of Bonn, Germany.

Statistical learning methods, such as Bayesian Networks, have gained a high popularity to infer cellular networks from high throughput experiments. However, the inherent noise in experimental data together with the typical low sample size limits their performance with high false positives and false negatives. Incorporating prior knowledge into the learning process has thus been identified as a way to address this problem, and principle a mechanism for doing so has been devised (Mukherjee & Speed, 2008). However, so far little attention has been paid to the fact that prior knowledge is typically distributed among multiple, heterogeneous knowledge sources (e.g. GO, KEGG, HPRD, etc.).
Here we propose two methods for constructing an informative network prior from multiple knowledge sources: Our first model is a latent factor model using Bayesian inference. Our second model is the Noisy-OR model, which assumes that the overall prior is a non-deterministic effect of participating information sources. Both models are compared to a naïve method, which assumes independence of knowledge sources. Extensive simulation studies on artificially created networks as well as full KEGG pathways reveal a significant improvement of both suggested methods compared to the naïve model. The performance of the latent factor model increases with larger network sizes, whereas for smaller networks the Noisy-OR model appears superior. Furthermore, we show that our informative priors significantly enhance the reconstruction accuracy of Bayesian

Network and Nested Effects Models. Finally, two examples, one from breast cancer and one from murine stem cell development highlight the utility of our approach.

Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization

Mehmet Gönen[1]
[1] Helsinki Institute for Information Technology, Finland

Motivation: Identifying interactions between drug compounds and target proteins has a great practical importance in the drug discovery process for known diseases. Existing databases contain very few experimentally validated drug-target interactions and formulating successful computational methods for predicting interactions remains challenging.
Results: In this study, we consider four different drug-target interaction networks from humans involving enzymes, ion channels, G-protein-coupled receptors, and nuclear receptors. We then propose a novel Bayesian formulation that combines dimensionality reduction, matrix factorization, and binary classification for predicting drug- target interaction networks using only chemical similarity between drug compounds and genomic similarity between target proteins. The novelty of our approach comes from the joint Bayesian formulation of projecting drug compounds and target proteins into a unified subspace using the similarities and estimating the interaction network in that subspace. We propose using a variational approximation in order to obtain an efficient inference scheme and give its detailed derivations. Lastly, we demonstrate the performance of our proposed method in three different scenarios: (a) exploratory data analysis using low-dimensional projections, (b) predicting interactions for the out-of-sample drug compounds, and (c) predicting unknown interactions of the given network.
Availability: Software and Supplementary Material are available at http://users.ics.tkk.fi/gonen/kbmf2k/
Contact: mehmet.gonen@aalto.fi

Targeted Retrieval of Gene Expression Measurements Using Regulatory Models

Elisabeth Georgii[1], Jarkko Salojärvi[2], Mikael Brosché[2], Jaakko Kangasjärvi[2] and Samuel Kaski[1]
[1] Helsinki Institute for Information Technology HIIT, Finland; [2] University of Helsinki at the plant stress research group, Finland

Motivation: Large public repositories of gene expression measurements offer the opportunity to position a new experiment into the context of earlier studies. While previous methods rely on experimental annotation or global similarity of expression profiles across genes or gene sets, we compare experiments by measuring similarity based on an unsupervised, data-driven regulatory model around pre-specified genes of interest. Our experiment retrieval approach is novel in two conceptual respects: (i) targetable focus and interpretability: the analysis is targeted at regulatory relationships of genes that are relevant to the analyst or come from prior knowledge; (ii) regulatory model-based similarity measure: related experiments are retrieved based on the strength of inferred regulatory links between genes.
Results: We learn a model for the regulation of specific genes from a data repository, and exploit it to construct a similarity metric for an information retrieval task. We use the Fisher kernel, a rigorous similarity measure that typically has been applied to utilize generative models in discriminative classifiers. Results on human and plant microarray collections indicate that our method is able to substantially improve the retrieval of related experiments against standard methods. Furthermore, it allows the user to interpret biological conditions in terms of changes in link activity patterns. Our study of the osmotic stress network for A. thaliana shows that the method successfully identifies relevant relationships around given key genes.
Availability: The code (R) will be available at http://research. ics.tkk.fi/mi/software.shtml at the time of publication.

Contact: elisabeth.georgii@aalto.fi, jarkko.salojarvi@helsinki.fi, samuel.kaski@hiit.fi

Metabolite identification and molecular fingerprint prediction via machine learning

Markus Heinonen[1], Huibin Shen[1], Nicola Zamboni[2] and Juho Rousu[1]
[1] Helsinki Institute for Information Technology, [2] Finland; ETH Zurich, Swizterland

Motivation: Metabolite identification from tandem mass spectra is an important problem in metabolomics, underpinning subsequent metabolic modelling and network analysis. Yet, currently this task requires matching the observed spectrum against a database of reference spectra originating from similar equipment and closely matching operating parameters, a condition that is rarely satisfied in public repositories. Furthermore, the computational support for identification of molecules not present in reference databases is lacking. Recent efforts in assembling large public mass spectral databases such as MassBank have opened the door for the development of a new genre of metabolite identification methods.
Results: We introduce a novel framework for prediction of molecular characteristics and identification of metabolites from tandem mass spectra using machine learning with the support vector machine (SVM). Our approach is to first predict a large set of molecular properties of the unknown metabolite from salient tandem mass spectral signals, and in the second step to use the predicted properties for matching against large molecule databases, such as PubChem. We demonstrate that several molecular properties can be predicted to high accuracy, and that they are useful in *de novo* metabolite identification, where the reference database does not contain any spectra of the same molecule.
Availability: An Matlab/Python package of the FingerID tool is freely available on the web at http://www.sourceforge.net/p/fingerid.
Contact: markus.heinonen@cs.helsinki.fi

Network Inference Using Steady-State Data and Goldbeter-Koshland Kinetics

Chris J Oates[1,2,3], Bryan Hennessey[4], Yiling Lu[5], Mills, Gordon B Mills[5], Sach Mukherjee[3]
[1] Centre for Complexity Science, University of Warwick, UK, [2] Department of Statistics, University of Warwick, UK, [3] Department of Biochemistry, Netherlands Cancer Institute, Amsterdam, The Netherlands, [4] Department of Medical Oncology, Beaumont Hospital, Dublin, Ireland, [5] Department of Systems Biology, The University of Texas M. D. Anderson Cancer Center, Houston, USA

Motivation: Network inference approaches are widely used to shed light on regulatory interplay between molecular players such as genes and proteins. Biochemical processes underlying networks of interest (e.g. gene regulatory or protein signalling networks) are gene- rally nonlinear. In many settings, knowledge is available concerning relevant chemical kinetics. However, existing network inference methods for continuous data are typically rooted in convenient statistical formulations which do not exploit chemical kinetics to guide inference.
Results: Here we present an approach to network inference for steady-state data that is rooted in nonlinear descriptions of biochemical mechanism. We use equilibrium analysis of chemical kinetics to obtain functional forms that are in turn used to infer networks using steady-state data. The approach we propose is directly applicable to conventional steady-state gene expression or proteomic data and does not require knowledge of either network topology or any kinetic parameters; both are simultaneously learned from data. We illustrate the approach in the context of protein phosphorylation networks, using data simulated from a recent mechanistic model and proteomic data from cancer cell lines. In the former, the true network is known and used for assessment, whilst in the latter results are compared against known biochemistry. We find that the proposed methodology is more effective at estimating network topology than methods based on linear models.
Availability: MATLAB R2009b code used to produce these results is provided in the Supplemental Information.
Contact: c.j.oates@warwick.ac.uk, s.mukherjee@nki.nl

Efficient Sampling for Bayesian Inference of Conjunctive Bayesian Networks

Thomas Sakoparnig[1] and Niko Beerenwinkel[2]
[1] Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, 4058 Basel, Switzerland, [2] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Motivation: Cancer development is driven by the accumulation of advantageous mutations and subsequent clonal expansion of cells harbouring these mutations, but the order in which mutations occur remains poorly understood. Advances in genome sequencing and the soon-arriving flood of cancer genome data produced by large cancer sequencing consortia hold the promise to elucidate cancer progression. However, new computational methods are needed to analyse these large datasets.
Results: We present a Bayesian inference scheme for Conjunctive Bayesian Networks, a probabilistic graphical model in which mutations accumulate according to partial order constraints and cancer genotypes are observed subject to measurement noise. We develop an efficient MCMC sampling scheme specifically designed to overcome local optima induced by dependency structures. We demonstrate the performance advantage of our sampler over traditional approaches on simulated data and show the advantages of adopting a Bayesian perspective when reanalysing cancer datasets and comparing our results to previous maximum likelihood-based approaches.
Availability: An R package including the sampler and examples is available at http://www.cbg.ethz.ch/software/bayes-cbn

Helios: discovering driver oncogenes

Felix Sanchez-Garcia[1], Bo-Juen Chen[1], Dylan Kotlier[1], Jose Silva[1] and Dana Pe'er[1].
[1] Columbia University, New York, United States

We present a novel unsupervised Bayesian integrative method for discovering driver genes in cancer. The algorithm incorporates signals from different data types such as copy number, gene expression, point mutations and functional assays into a single probabilistic score. The method was applied to breast cancer, correctly identifying known drivers and uncovering novel oncogenes that were validated both in vitro and in vivo.

Poster Presentations

1. Askar Obulkasim, Gerrit Meijer and Mark van de Wiel. *Optimal Cut Finding for the Hierarchical Clustering Using Background Information*
2. Fazel Famili, Ziying Liu and Sieu Phan. *Identifying meaningful patterns in gene expression time-series data*
3. Karin Proell, Witold Jacak and Michelle Epstein. *Unsupervised Neural Networks based Feature Selection for Biomarker Analysis*
4. Jana Kludas, Mikko Arvas and Juho Rousu. *Data Fusion and Feature Selection for Supervised Protein Interaction Prediction*
5. Jia-Hong Wu, Yun-Ru Sun and Grace Shieh. *Inferring co-regulation of transcription factors and microRNAs in human breast cancer*
6. Yongsoo Kim, Sehyun Chae, Daehee Hwang and Seungjin Choi. *Probabilistic Model for Protein Quantification using Mass Spectrometry*
7. Jong Kyoung Kim, Yongsoo Kim and Seungjin Choi. *Infinite Mixture Model for Inferring Transcription Factor Binding Mechanisms*
8. Witold Dyrka, Jean-Christophe Nebel and Malgorzata Kotulska. *Structural annotation of proteins using probabilistic context-free grammars*
9. Je-Keun Rhee, Soo-Jin Kim, Hyo-Jeong Ban, Woongchang Yoon, Kiejung Park and Byoung-Tak Zhang. *Systems genetics analysis through identification of multiple SNP interaction by evolutionary learning*
10. Waqar Ali and John Pinney. *Ascertainment bias in protein interaction datasets.*
11. Neetika Nath, John B. O. Mitchell and Lazaros Mavridis. *Global Analysis of Enzyme Reaction Mechanisms*
12. S. M. Minhaz Ud-Dean and Rudiyanto Gunawan. *TRACE: Transitive Reduction and Closure Estimation of Gene Regulatory Networks*
13. Sanja Brdar, Vladimir Crnojević and Blaž Zupan. *Non-Negative Matrix Factorization as a Framework for Integration of Gene Clusters*
14. James Barrett and Ton Coolen. *Dimensionality Reduction and Data Integration using a Bayesian Latent Variable Model*
15. Leo Lahti. *Bayesian online-learning method for scalable preprocessing of microarray atlases*
16. Zhanpan Zhang, Xinping Cui, Daniel Jeske and James Borneman. *Co-clustering Scatter Plots Using Data Depth Measures*

MLSB Program Committee

Hendrik Blockeel	K.U. Leuven, Belgium
Saso Dzeroski	Jozef Stefan Institute, Slovenia
Pierre Geurts	University of Liege, Belgium
Lars Kaderali	University of Technology Dresden, Germany
Ross King	University of Wales, Aberystwyth, UK
Stefan Kramer	University of Mainz, Germany
Yves Moreau	K.U. Leuven, ESAT-SCD, Belgium
Sach Mukherjee	University of Warwick, UK
Uwe Ohler	Duke University, USA
Dana Peer	Columbia Univeristy, USA
John Pinney	Imperial College London, UK
Simon Rogers	University of Glasgow, UK
Juho Rousu	Aalto University, Finnland
Celine Rouveirol	LIPN, Université Paris 13, France
Yvan Saeys	Ghent University, Belgium
Peter Sykacek	Boku University, Austria
Ljupco Todorovski	University of Ljubljana, Slovenia
Achim Tresch	Max Planck Institute for Plant Breeding Research
Koji Tsuda	National Institute of Advanced Industrial Science and Technology, Japan
Jean-Philippe Vert	Ecole des Mines de Paris, France
Filip Zelezny	Czech Technical University, Czech Republic

OUP Bioinformatics Virtual Issue Reviewers

Chloé-Agathe Azencott	MPI for Intelligent Systems & MPI for Developmental Biology, Germany
Jonas Behr	MSKCC, New York, USA
Tim Beissbarth	University of Göttingen, Germany
Hendrik Blockeel	K.U. Leuven, Belgium
Florence d'Alché-Buc	Université d'Evry, France
Philipp Drewe	MSKCC, New York, USA
Saso Dzeroski	Jozef Stefan Institute, Slovenia
Dave duVerle	Kyoto University Bioinformatics Center, Japan
Alvaro González	MSKCC, New York, USA
Dirk Husmeier	JCMB, Edinburgh, UK
Lars Kaderali	Dresden University of Technology, Germany
Andre Kahles	MSKCC, New York, USA
Stefan Kramer	University of Mainz, Germany
Xinghua Lou	MSKCC, New York, USA
Limin Li	Xi'an University, China
Christoph Lippert	Microsoft Research L.A., USA
Elena Marchiori	Radboud University, Netherlands
Anirban Mukhopadhyay	University of Kalyani, India
Richard Neher	MPI for Developmental Biology, Germany
Raphael Pelosoff	MSKCC, New York, USA
John Pinney	Imperial College London, UK
Barbara Rakitsch	MPI for Intelligent Systems & MPI for Developmental Biology, Germany
Matthias Rarey	Zentrum für Bioinformatik, Hamburg, Germany
Boris Reva	MSKCC, New York, USA
Simon Rogers	University of Glasgow, UK
Juho Rousu	Aalto University, Finnland
Céline Rouveirol	LIPN, Université Paris 13, France
Yvan Saeys	Ghent University, Belgium
Ram Samudrala	University of Washington, Seattle, USA
Guido Sanguinetti	University of Edinburgh, UK
Ron Shamir	The Blavatnik School of Computer Science, Israel
Hyunjung Shin	Ajou University, Korea
Henry Soldano	Université Paris-Nord, France
Oliver Stegle	MPI for Intelligent Systems & MPI for Developmental Biology, Germany
Richard Stein	Princeton University, USA

Peter Sykacek	Boku University, Austria
Koji Tsuda	AIST Tokyo, Japan
Jean-Philippe Vert	Ecole des Mines de Paris, France
Louis Wehenkel	University of Liège, Belgium
Christian Widmer	MSKCC, New York, USA
Michael Zhang	Cold Spring Harbor, USA