MLSB 13

The seventh workshop on Machine Learning in Systems Biology

Berlin, July 19-20, 2013

A special interest group meeting at the 21st annual international conference on Intelligent Systems in Molecular Biology (ISMB)

MLSB13 Chairs

Uwe Ohler	Berlin Institute for Molecular Systems Biology / Max Delbrueck Center, Germany
Jean-Philippe Vert	Institut Curie / Mines ParisTech, France

Scientific Program Committee

Karsten Borgwardt (Max Planck Institute, Tuebingin)
Florence d'Alché-Buc (University of Evry, France)
Sašo Džeroski (Jožef Stefan Institute, Slovenia)
Paolo Frasconi (University of Florence, Italy)
Pierre Geurts (University of Liège, Belgium)
Lars Kaderali (TU Dresden, Germany)
Samuel Kaski (Aalto University and University of Helsinki, Finland)
Ross King (Manchester University, UK)
Stefan Kramer (University of Mainz, Germany)
David Kreil (University of Warwick, UK)
Christina Leslie (Memorial Sloan-Kettering Cancer Center, USA)
Yves Moreau (Katholieke Universiteit Leuven, Belgium)
Mahesan Niranjan (University of Southampton, UK)
John Pinney (Imperial College London , UK)
Magnus Rattray (Manchester University, UK)
Simon Rogers (University of Glasgow, UK)
Juho Rousu (University of Helsinki, Finland)
Céline Rouveirol (Paris 13 University, France)
Yvan Saeys (University of Gent, Belgium)
Guido Sanguinetti (University of Edinburgh, UK)
Johannes Soeding (Ludwig-Maximilian-University Munich, Germany)
Peter Sykacek (BOKU University, Austria)
Ljupco Todorovski (University of Ljubljana, Slovenia)
Achim Tresch (MPI for Plant Breeding, Cologne)
Koji Tsuda (National Institute of Advanced Industrial Science and Technology, Japan)
Louis Wehenkel University of Liège, Belgium)
Filip Zelezny (Czech Technical University in Prague, Czech Republic)

Program Friday July 19th

08:45-09:45	**Invited speaker: Guido Sanguinetti.** *Hybrid stochastic models of gene expression.*
09:45-10:15	**Diane Oyen, Alexandru Niculescu-Mizil, Rachel Ostroff and Alex Stewart.** *Controlling the Precision-Recall Tradeoff in Differential Dependency Network Analysis.*
10:15-10:45	Coffee break
10:45-11:15	**Carlo Vittorio Cannistraci, Gregorio Alanis-Lobato and Timothy Ravasi.** *From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks.*
11:15-11:45	**Joris Mooij and Tom Heskes.** *Discovering Cyclic Causal Models from Continuous Equilibrium Data.*
11:45-12:15	**Markus Heinonen, Olivier Guipaud, Fabien Milliat, Valerie Buard, Beatrice Micheau and Florence D'Alche-Buc.** *Time-dependent gaussian process regression and significance analysis for sparse time-series.*
12:15-13:30	Lunch
13:30-14:30	**Invited speaker: Lani Wu.** *Analysis of heterogeneity can reveal network motifs.*
14:30-15:00	**Marco Frasca, Alberto Bertoni and Giorgio Valentini.** *An unbalance-aware network integration method for gene function prediction.*
15:00-16:30	Poster session (coffee from 15:30-16:00)
16:30-17:00	**Fayyaz-Ul-Amir Afsar, Brian Geiss and Asa Ben-Hur.** *PAIRpred: Prediction of partner-specific interacting residues from sequence and structure.*
17:00-17:30	**Paula Petrone, Meir Glick, Jeremy Jenkins, Paul Selzer, Benjamin Simms, Anne Mai Wassermann, Florian Nigsch, Eugen Lounkine, Peter Kutchukian, Allen Cornett, Zhang Deng and John W Davies.** *Navigating chemical-biology space: comparing and selecting compounds based on biological activity.*
17:30-18:00	**Shankar Vembu and Quaid Morris.** *An Efficient Algorithm to Integrate Network and Attribute Data for Gene Function Prediction.*

Program Saturday July 20th

08:45-09:45	**Invited speaker: Sayan Mukherjee.** *Modeling quantitative phenotypes.*
09:45-10:15	**Xin Wang, Ke Yuan, Christoph Hellmayr, Wei Liu and Florian Markowetz.** *Capturing rewiring events during network evolution underlying dynamic biological processes.*
10:15-10:45	Coffee break
10:45-11:15	**Elena Rivas and Sean Eddy.** *A probabilistic evolutionary model compatible with standard affine gap cost sequence alignment.*
11:15-11:45	**Néhémy Lim, Yasin Senbabaoglu, George Michailidis and Florence D'Alché-Buc.** *Nonparametric modeling for gene regulatory network inference using boosting and operator-valued kernels.*
11:45-12:15	**Johannes Stephan, Oliver Stegle and Andreas Beyer.** *Mixed Random Forests – non-linear feature selection and response prediction in structured populations.*
12:15-13:30	Lunch & poster viewing
13:30-14:00	**Yawwani Gunawardana and Mahesan Niranjan.** *Bridging the Gap Between Transcriptome and Proteome Measurements Identifies Post Translationally Regulated Genes.*
14:00-14:30	**Nataly Maimari, Calin-Rares Turliuc, Krysia Broda, Antonis Kakas, Rob Krams and Alessandra Russo.** *ARNI: Abductive inference of complex regulatory network structure.*
14:30-15:00	**Ciira Maina, Filomena Matarese, Korbinian Grote, Hendrik Stunnenberg, George Reid, Antti Honkela, Neil Lawrence and Magnus Rattray.** *A Probabilistic Model of Transcription Dynamics applied to Estrogen Signalling in Breast Cancer Cells.*
15:00-15:30	**Steven M. Hill, Nicole K. Nesser, Paul T. Spellman and Sach Mukherjee.** *Data-driven inference of causal molecular networks and systematic validation of inference performance.*
15:30-16:00	Coffee break
16:00-17:00	**Invited speaker: John Marioni.** *Single-cell RNA sequencing -- challenges and opportunities.*
17:00	Closing remarks

Introduction

The aim of this workshop is to contribute to the cross-fertilization between the research in machine learning methods and their applications to systems biology (i.e., complex biological and medical questions) by bringing together method developers and experimentalists. We encourage submissions bringing forward methods for discovering complex structures (e.g. interaction networks, molecule structures) and methods supporting genome-wide data analysis.

Berlin is the seventh in the series, which has since 2011 been organized as a workshop/special interest group at the European Conference on Computational Biology (ECCB)/Intelligent Systems in Molecular Biology (ISMB). 16 platform presentations were selected from over 40 submissions; four invited speakers present work spanning statistical genetics, regulatory and signaling networks, and single cell expression dynamics.

We are thankful to our sponsors, and we thank all those involved for their efforts, in particular Steven Leard in the ISMB conference HQ. All of you has made it possible to put together an exciting meeting. We hope you will enjoy it as much as we will!

Jean-Philippe & Uwe

Sponsors

The Berlin Institute for Medical Systems Biology has provided support for invited speakers.

F1000 Research has provided prizes for the best poster/platform presentations.

Abstracts of platform presentations

[in order of the workshop schedule]

Hybrid stochastic models of gene expression

Guido Sanguinetti

University of Edinburgh, UK
gsanguin@inf.ed.ac.uk

The dynamics of gene expression is often modelled through systems of nonlinear Ordinary Differential Equations (ODEs). While these provide a powerful and flexible framework which can leverage centuries of mathematical research, identification of nonlinear ODEs from data presents formidable (and possibly unsurmountable) statistical challenges. I will describe a novel approach which abstracts the complex dynamics of gene regulatory networks using simple building blocks: promoters (regulatory elements upstream of genes), modelled as telegraph processes, and proteins, modelled as conditionally Gaussian processes (conditioned on promoters). I will describe a novel approximate inference framework for this class of models that yields scalable and accurate posterior estimation, and illustrate the method on a study of the structure of the circadian oscillator in the picoalga Ostreococcus tauri.

Joint work with Andrea Ocone and (earlier part) Manfred Opper.

Controlling the Precision-Recall Tradeoff in Differential Dependency Network Analysis

Diane Oyen doyen@cs.unm.edu
University of New Mexico, Albuquerque, NM USA

Alexandru Niculescu-Mizil alex@nec-labs.com
NEC Laboratories, Princeton, NJ USA

Rachel Ostroff, Alex Stewart
{rostroff,astewart}@somalogic.com
SomaLogic Inc., Boulder, CO USA

Graphical models have gained a lot of attention recently for representing dependencies among variables in multivariate data. Often, domain scientists are looking specifically for differences among related dependency networks, such as for regulatory networks of different species, or for diseased versus healthy populations. The standard method for finding these differences is to learn individual networks independently and then compare them. We show that this approach is prone to high false discovery rates (low precision) that can render the analysis useless. We then show how techniques developed in the transfer learning literature can be used to obtain acceptable false discovery rates, at the cost of finding a reduced number of differences. Transfer learning techniques provide a natural mechanism to smoothly adjust this precision-recall tradeoff to cater to the requirements of the analysis conducted. We present two oncological case studies where these techniques are used to extract useful differential networks that shed light on biological processes involved in cancer.

From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks

Carlo Vittorio Cannistraci[1, 2, †, *], Gregorio Alanis-Lobato[1, 2, †] and Timothy Ravasi [1, 2, *]

[1]Integrative Systems Biology Laboratory, Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Ibn Al Haytham Bldg. 2, Level 4, Thuwal 23955-6900, Kingdom of Saudi Arabia
[2]Division of Medical Genetics, Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 USA.
[†] These authors contributed equally to this work
[*]corresponding authors: kalokagathos.agon@gmail.com and timothy.ravasi@kaust.edu.sa

Growth and remodelling impact the network topology of complex systems, yet a general theory explaining how new links arise between existing nodes has been lacking, and little is known about the topological properties that facilitate link-prediction.
Here we investigate the extent to which the connectivity evolution of a network might be predicted by mere topological features. The results we obtained in simulations of link prediction in several complex networks of different type and size (brain connectomes, protein interactomes (PPI), social and ecological networks) show how the link/community-based strategy adopted by our method triggers substantial prediction improvements, because it accounts for the singular topology of several real networks organised in multiple local communities - a tendency that we named local-community-paradigm (LCP).
We observe that LCP networks are mainly formed by weak interactions and characterise heterogeneous and dynamic systems that use self-organisation as a major adaptation strategy. These systems seem designed for global delivery of information and processing via multiple local modules. Conversely, non-LCP networks have steady architectures formed by strong interactions, and seem designed for systems in which information/energy storage is crucial.

http://www.nature.com/srep/2013/130408/srep01613/full/srep01613.html

Discovering Cyclic Causal Models from Continuous Equilibrium Data

Joris M. Mooij[1,2], Tom Heskes[2]

1Informatics Institute, University of Amsterdam, the Netherlands
2Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands

We propose a method for learning cyclic causal models from a combination of observational and interventional equilibrium data. Novel aspects of the proposed method are its ability to model continuous data with nonlinear relationships and to deal with feedback loops. Within the context of biochemical reactions, we also propose a novel way of modeling interventions that modify the activity of compounds instead of their abundance. We apply the method to reconstruct a cellular signaling network from the flow cytometry data measured by Sachs et al. (2005). We find evidence in the data for feedback loops and obtain a more accurate quantitative description of the data at comparable model complexity.

This abstract describes the work reported in (Mooij and Heskes, 2013).

Time-dependent gaussian process regression and signi_cance analysis for sparse time-series

Markus Heinonen [1,2†], Olivier Guipaud [3], Fabien Milliat [3], Valerie Buard [3], Beatrice Micheau [3] and Florence d'Alche-Buc [1,2]

[1]IBISC, Universite d''Evry Val d'Essonne, France
[2]LRI, INRIA-Saclay, Universite Paris Sud, France
[3]LRTE, Institut de Radioprotection et de Surete Nucleaire (IRSN), France
[†]markus.heinonen@ibisc.fr

Gaussian process regression (GPR) has been extensively used for modelling and differential testing of biological time-series measurements due to its robustness and interpretability. However, the standard gaussian process assumes stationary model dynamics and is a poor fit for common perturbation experiments, where we expect to see rapid changes after the perturbation and diminishing rate of state change as the cell returns back to a stable state.

A common application of time-series measurements is the testing of significant difference between two time-serie profiles. The currently used two-sample differential tests, based on gaussian processes, focus on comparing model likelihoods over a subset of measured time-points, and hence necessitate dense measurements to cover the time axis.

We address these problems by proposing time-dependent extensions to both gaussian process regression and significance analysis between time-series. We propose a time-dependent noise model and time-dependent covariance priors, suitable for perturbation experiments. We utilise a novel model inference criteria for sparse measurements, which results in more informative models along time. We propose two novel differential tests for time-series, that both allow significance testing at non-observed time-points. We apply the extended GPR model for analysis of differential expression of irradiated human umbilical vein endothelial cell (HUVEC) transcriptomics dataset.

Analysis of cellular heterogeneity can reveal network motifs.

Lani Wu

University of Texas Southwestern Medical Center, Dallas TX, USA
lani.wu@utsouthwestern.edu

Neutrophil polarity relies on local, mutual inhibition to restrict incompatible signaling circuits to the leading and trailing edges. Intuitively, mutual inhibition alone would lead to populations of cells having strong fronts and weak backs or vice versa, which has not been observed experimentally. Analysis of cell-to-cell variation in images of polarized primary human neutrophils revealed that back polarity remains consistent despite changes in front strength. How is this buffering achieved? Pharmacological perturbations revealed a new functional role for microtubules to buffer back polarity by mediating positive, long-range crosstalk from front to back. Further, a systematic computational search over network topologies found that a long-range, positive link from front to back is necessary for back buffering.

An unbalance-aware network integration method for gene function prediction

Marco Frasca, Alberto Bertoni and Giorgio Valentini

Dipartimento di Informatica
Universita degli Studi di Milano, Italy
{frasca, bertoni,valentini}@di.unimi.it

Data integration and the unbalance between functionally annotated and unannotated genes are relevant items in the context of network-based gene function prediction. Even if both these topics have been analyzed in recent works, to our knowledge no network integration methods, specific for unbalanced functional classes have been proposed in this context. We introduce an unbalance-aware network integration method based on the recently proposed COSNet algorithm, and we apply it to the genome-wide prediction of Gene Ontology terms with the M. musculus model organism.

PAIRpred: Prediction of partner-specific interacting residues from sequence and structure

Fayyaz-ul-Amir Afsar Minhas[1,*] Brian J. Geiss[2] and Asa Ben-Hur[3,*]

[1,3]Department of Computer Science, [2]Department of Microbiology, Immunology and Pathology
Colorado State University, Fort Collins, Colorado 80523, USA.
[1]fayyazafsar@gmail.com, [2]Brian.Geiss@colostate.edu, [3]asa@cs.colostate.edu

Motivation: Computational prediction of whether a pair of residues belonging to two different proteins in a complex interact with each other is a relatively unexplored problem and the accuracy of existing methods in this area is very low.
Results: We present a partner-specific protein-protein interaction site prediction method we call PAIRpred that uses either sequence information alone or a combination of sequence and structure fea-tures. The proposed method offers state of the art accuracy, and we find that incorporating features computed from structure provides a big improvement in accuracy. We have found that performance de-creases with the degree of conformational change of a complex upon binding. We demonstrate our method using a case-study of a human-virus protein interaction.
Availability: Python code and data files for PAIRpred are available at http://combi.cs.colostate.edu/supplements/pairpred

Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity

Paula M. Petrone,† Benjamin Simms,† Florian Nigsch, Eugen Lounkine, Peter Kutchukian, Allen Cornett, Zhan Deng, John W. Davies, Jeremy L. Jenkins and Meir Glick

Center for Proteomic Chemistry, Novartis Institutes for Biomedical Research Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

Since the advent of high-throughput screening (HTS), there has been an urgent need for methods that facilitate the interrogation of large-scale chemical biology data to build a mode of action (MoA) hypothesis. This can be done either prior to the HTS by subset design of compounds with known MoA or post HTS by data annotation and mining. To enable this process, we developed a tool that compares compounds solely on the basis of their bioactivity: the chemical biological descriptor “high-throughput screening fingerprint” (HTS-FP). In the current embodiment, data are aggregated from 195 biochemical and cell-based assays developed at Novartis and can be used to identify bioactivity relationships among the in-house collection comprising 1.5 million compounds. We demonstrate the value of the HTS-FP for virtual screening and in particular scaffold hopping. HTS-FP outperforms state of the art methods in several aspects, retrieving bioactive compounds with remarkable chemical dissimilarity to a probe structure. We also apply HTS-FP for the design of screening subsets in HTS. Using retrospective data, we show that a biodiverse selection of plates performs significantly better than a chemically diverse selection of plates, both in terms of number of hits and diversity of chemotypes retrieved. In this context, we demonstrate that rather than chemical diversity, the biological diversity of a compound library is an essential requirement for hit identification.

An Efficient Algorithm to Integrate Network and Attribute Data for Gene Function Prediction

Shankar Vembu and Quaid Morris

University of Toronto, Canada
fshankar.vembu,quaid.morrisg@utoronto.ca

Several graph-based machine learning algorithms have been proposed to combine multiple functional interaction networks for gene function prediction. However, these algorithms cannot be used to integrate feature-based data sources with networks. We propose an efficient learning algorithm to integrate heterogeneous data sources, including functional interaction networks and feature-based data sources, for gene function prediction. Our method, LMGraph, consists of two steps. In the first step, we extract a small set of discriminative features from the nodes of graphs. In the second step, we apply a simple weighting scheme in conjunction with linear classifiers to combine these features. This two-step procedure allows us to (i) learn highly scalable and computationally efficient linear classifiers, (ii) and seamlessly combine feature-based data sources with networks. Experiments on multiple functional interaction networks from three species (mouse, fly, c.elegans) with tens of thousands of nodes and hundreds of gene ontology biological process categories demonstrate the efficacy of our method.

Modeling Quantitative Phenotypes

Sayan Mukherjee

Department of Statistical Sciences, Duke University, Durham NC, USA
sayan@stat.duke.edu

In this talk we consider two problems in modeling quantitative phenotypes.
The first problem is estimating the genetic covariance matrix(G-matrix) of high-dimensional traits. The second problem involves measuring distances between bones (2-dimension surfaces embedding in 3-dimensions).

Quantitative genetic studies that model complex, multivariate phenotypes are important for both evolutionary prediction and artificial selection. For example, changes in gene expression can provide insight into developmental and physiological mechanisms that link genotype and phenotype. However, classical analytical techniques are poorly suited to quantitative genetic studies of gene expression where the number of traits assayed per individual can reach many thousand. Here, we derive a Bayesian genetic sparse factor model for estimating the genetic covariance matrix (G-matrix) of high-dimensional traits, such as gene expression, in a mixed effects model.

A method to measure distances between surfaces, such as bones, when the surfaces are qualitatively different, for example they are not isomorphic. The motivation for estimating these distances is to better understand selective pressure. The method uses ideas from computational topology and places them in a probabilistic framework.

Capturing rewiring events during network evolution underlying dynamic biological processes

Xin Wang, Ke Yuan, Christoph Hellmayr, Wei Liu, Florian Markowetz

Cancer Research UK Cambridge Institute, University of Cambridge
{xin.wang;ke.yuan;florian.markowetz}@cruk.cam.ac.uk

We propose hidden Markov nested effects models (HM-NEMs) to reconstruct evolving signalling networks from indirect effects of systematic perturbations over time. A key strength of HM-NEMs is that the MCMC sampling algorithm we developed is able to infer the most probable time-varying network while estimating the parameter quantifying the intrinsic feature of network evolution. With two real world biological applications, we demonstrate how HM-NEMs gain insights to the mechanism of network evolution underlying complex dynamic systems.

A probabilistic evolutionary model compatible with standard affine gap cost sequence alignment

Elena Rivas and Sean R. Eddy
Janelia Farm Research Campus, Howard Hughes Medical Institute,
Ashburn VA 20147, USA
rivase@janelia.hhmi.org

Inference of sequence homology is inherently an evolutionary question, dependent upon evolutionary divergence times. However, the insertion and deletion penalties in the most widely used methods for inferring homology by sequence alignment, including BLAST and profile hidden Markov models (profile HMMs), are not based on any explicitly time-dependent evolutionary model. Previous evolutionary models of insertion and deletion events have either been for simpler parameterizations (linear-cost, as opposed to affine-cost), or yield solutions incompatible with standard sequence comparison, or have not been analytically solved.

Here we have solved a model of sequence evolution that provides analytical expressions that describe the occurrence of insertion and deletion events over time, in a fashion compatible with the parameterization of standard profile HMM methods, and by extension, other standard sequence alignment methods that use affine gap costs. This is a birth-death model that at any given time allows infinitesimal events that independently insert or delete a whole new insert of several residues or one single residue within an existing insert. By being probabilistic and allowing affine insertions and deletions, this new evolutionary model can be directly applied to standard profile and pair hidden Markov models (HMMs) used for protein and DNA homology searches. Under an evolutionary interpretation, each Match/Delete/Insert ``unit'' of the HMM describes the possible fates of an ancestral residue (whether substituted or deleted, and possibly sustaining an arbitrary number of insertions before the next ancestral unit). Under our evolutionary parameterization, the transition probabilities of the HMM become analytic time-dependent functions that depend on position-specific (``profiled'') rate parameters.

We have also solved evolutionary models corresponding to more complicated HMM architectures that permit a more realistic treatment of insertions than just affine. Obtaining an exact evolutionary model that fits naturally into standard profile HMM models provides us with a direct and principled means of integrating phylogenetic modeling into most existing methods for pairwise and profile sequence alignment. This result smooths the way for the integration of phylogenetic modeling into standard methods for sequence alignment and homology detection based on profile and pair HMMs.

For a given profile HMM parameterized as a model of a particular alignment, one can construct an evolutionary version of the HMM where its position-specific rates can be estimated under certain assumptions without need for any additional information. At this meeting, we will provide a comparison of the power of detecting remote homology by the

original and the time-dependent version of profile HMMs implemented by HMMER.

Nonparametric modeling for gene regulatory network inference using boosting and operator-valued kernels

Nehemy Lim
IBISC EA 4526 Universite d' Evry-Val d'Essonne
23 Bd de France, 91000, Evry, France
nlim@ibisc.univ-evry.fr

Yasin Senbabaoglu
Department of Computational Medicine & Bioinformatics
University of Michigan Ann Arbor, MI 48109-2218
yasinsen@umich.edu

George Michailidis
Department of Statistics,
University of Michigan Ann Arbor, MI 48109-1107
gmichail@umich.edu

Florence d'Alche-Buc
INRIA-Saclay, LRI umr CNRS 8623 Universite Paris Sud, France
IBISC EA 4526, Universite d' Evry-Val d'Essonne
23 Bd de France, 91000, Evry, France
florence.dalche@ibisc.fr

Reverse engineering of gene regulatory networks remains a central challenge in computational systems biology, despite recent advances facilitated by benchmark in-silico challenges that have aided in calibrating their performance. A number of approaches using either perturbation (knock-out) or wild-type time series data have appeared in the literature addressing this problem, with the latter employing linear temporal models.
Nonlinear dynamical models are particularly appropriate for this inference task given the generation mechanism of the time series data. In this study, we introduce a novel nonlinear autoregressive model based on operator-valued kernels that simultaneously learns the model parameters, as well as the network structure. A flexible boosting algorithm (OKVAR-Boost) that shares features from L2-boosting and randomization-based algorithms is developed to perform the tasks of parameter learning and network inference for the proposed model. Specifically, at each boosting iteration, a regularized operator-valued kernel based vector autoregressive model (OKVAR) is trained on a random subnetwork. The final model consists of an ensemble of such models. The empirical estimation of the ensemble model's Jacobian matrix provides an estimation of the network structure. The performance of the proposed algorithm is evaluated on a number of benchmark data sets from the DREAM3 challenge. The high quality results obtained strongly indicate that it outperforms existing approaches.

Mixed Random Forests - non-linear feature selection and response prediction in structured populations

Johannes Stephan, Oliver Stegle*, Andreas Beyer

BIOTEC TU-Dresden, Germany
*EBI Hinxton, UK
andreas.beyer@biotec.tu-dresden.de

Reliable identification of causally relevant genetic features such as single nucleotide polimorphysms (SNPs) is a critical yet still unsolved task. The true functional dependencies that explain phenotypic variance are often complex, usually involving large numbers of genetic features that have both linear additive and non-linear effects. In addition, individual features are frequently confounded by population structure.

Here, we propose a random bagging method to address the challenge of selecting non-linear genetic features while accounting for population structure. Our model combines the advantages of random forests, identifying non-linear effects in an efficient manner, with mixed modeling strategies to handle and account for confounding. Benchmarking on large-scale applications in statistical genetics, we find that the proposed mixed random forest outperforms several state-of-the-art methods not only in the selection of linear additive and non-linear features, but also in terms of predictive performance.

Bridging the Gap Between Transcriptome and Proteome Measurements Identifies Post Translationally Regulated Genes

Yawwani Gunawardana and Mahesan Niranjan

School of Electronics and Computer Science, University of Southampton, Southampton, UK
mn@ecs.soton.ac.uk

Despite much dynamical cellular behaviour being achieved by accurate regulation of protein concentrations, messenger RNA abundances, measured by microarray technology, and more recently by deep sequencing techniques are widely used as proxies for protein measurements. While for some species and under some conditions, there is good correlation between transcriptome and proteome level measurements, such correlation is by no means universal due to post-transcriptional and post-translational regulation, both of which are highly prevalent in cells. Here, we seek to develop a data-driven machine learning approach to bridging the gap between these two levels of high throughput omic measurements on Saccharomyces cerevisiae and deploy the model in a novel way to uncover mRNA-protein pairs that are candidates for post-translational regulation. The application of feature selection by sparsity inducing regression (l_1 norm regularization) leads to a stable set of features; i.e. mRNA, ribosomal occupancy, ribosome density, tRNA adaptation index (tAI) and codon bias, while achieving a feature reduction from 37 to five. L_1 norm regularization predictor employed with these features is capable of predicting protein concentrations fairly accurately (R^2 = 0.86). Proteins whose concentration cannot be predicted accurately, taken as outliers with respect to the L_1 norm regularization predictor, are shown to have annotation evidence of post-ranslational modification, significantly more than random subsets of similar size $p < 0.05$. In a data mining sense, this work also shows a wider point that outliers with respect to a learning method can carry meaningful information about a problem domain.

ARNI: Abductive inference of complex regulatory network structure

Nataly Maimari[1,2], Calin-Rares Turliuc[2], Krysia Broda[2], Antonis Kakas[3], Rob Krams[1], and Alessandra Russo[2]

1. Dept. of Bioengineering, Imperial College London, UK
2. Dept. of Computing, Imperial College London, UK
3. Dept. of Computer Science, University of Cyprus

A fundamental challenge in systems biology is the extraction of integrated signalling-transcriptional networks from high throughput datasets. Current inference methods have limited applicability, relying on cause-effect pairs or systematically perturbed datasets and fail to capture complex network structures such as feedback loops. Here we present a novel framework, ARNI, based on abductive inference, that addresses these limitations. Using a network controlling T-cell differentiation, we illustrate that ARNI can effectively capture complex structures not detected by any of the other methods. In addition, we demonstrate how ARNI, integrating elements of model checking and state prediction, can act as a scientific assistant to help experts explore biological hypotheses. The method presented in this work achieves wider applicability and improved expressiveness

A Probabilistic Model of Transcription Dynamics applied to Estrogen Signalling in Breast Cancer Cells

Ciira wa Maina1, Filomena Matarese2, Korbinian Grote3, Hendrik G. Stunnenberg 2, George Reid4, Antti Honkela5, Neil D. Lawrence1, Magnus Rattray6

1 University of Sheffield, UK
2 Radboud University, NL
3 Genomatix Software GmbH
4 Institute for Molecular Biology, Mainz, Germany
5 University of Helsinki, Finland
6 University of Manchester, UK
E-mail: c.maina@sheffield.ac.uk, magnus.rattray@manchester.ac.uk, n.lawrence@sheffield.ac.uk

Gene transcription mediated by RNA polymerase II (pol-II) is a key step in gene expression. The dynamics of pol-II moving along the transcribed region influences the rate and timing of gene expression.

In this work we present a probabilistic model of transcription dynamics which is fitted to pol-II occupancy time course data measured using ChIP-Seq. The model can be used to estimate transcription speed and to infer the temporal pol-II activity profile at the gene promoter. Model parameters are determined using either maximum likelihood estimation or via Bayesian inference using Markov chain Monte Carlo sampling. The Bayesian approach provides confidence intervals for parameter estimates and allows the use of priors that capture domain knowledge, e.g. the expected range of transcription speeds, based on previous experiments. The model describes the movement of pol-II down the gene body and can be used to identify the time of induction for transcriptionally engaged genes. By clustering the inferred promoter activity time profiles, we are able to determine which genes respond quickly to stimuli and group genes that share activity profiles and may therefore be co-regulated.

We apply our methodology to biological data obtained using ChIP-seq to measure pol-II occupancy genome-wide when MCF-7 human breast cancer cells are treated with estradiol (E2). The transcription speeds we obtain agree with those obtained previously for smaller numbers of genes with the advantage that our approach can be applied genome-wide. We validate the biological significance of the pol-II promoter activity clusters by investigating cluster-specific transcription factor binding patterns and determining canonical pathway enrichment.

Data-driven inference of causal molecular networks and systematic validation of inference performance

Steven M .Hill[1], Nicole K. Nesser[2], Paul T. Spellman[2], Sach Mukherjee[1]

1 The Netherlands Cancer Institute, Amsterdam, The Netherlands
2 Oregon Health and Science University, Portland, OR 97239

The elucidation of molecular network topology from time series data remains a challenging problem and continues to be an active area of research. It is often of interest to infer networks in which directed edges can be interpreted as causal relationships between molecular components. Such causal relationships play an important role in the development and progression of diseases, such as cancer. Therefore the ability to infer context-specific causal networks in a data-driven manner is important for better understanding disease mechanisms and response to therapy. However, this requires experimental interrogation of multiple components through time, together with robust and scalable statistical approaches for causal network estimation.

In this work we focus on protein signalling networks in cancer and utilise directed graphical models known as dynamic Bayesian networks (DBNs) to make inferences regarding network structure from breast cancer proteomic data. A Bayesian framework is used, enabling the incorporation of existing biology via an informative prior distribution over networks. Instead of applying approximate schemes, such as MCMC, we exploit a connection between variable selection and network inference to enable exact calculation of posterior probabilities of interest. This results in a scalable procedure which is computationally efficient and requires no MCMC convergence diagnostics. We also show how the DBN approach can be extended to model data in which certain nodes are intervened upon. Such interventional data can aid causal inference, but our results demonstrate that it is vital to take the interventions into account in a suitable manner.

Network inference procedures are often assessed using simulated data, where a ‘gold-standard’ network structure is available. For real-world systems, lack of such a ‘gold-standard’ makes assessment of performance challenging. Results are often compared to the literature or small subsets of inferred edges are validated in independent experiments. We outline an approach that uses interventional data to enable systematic experimental validation of estimated causal networks. We find that our interventional DBN approach is able to recover causal network structure in this challenging mammalian signalling setting.

The data presented in this work forms part of the 2013 DREAM8 network inference challenge.

Single-cell RNA-sequencing - challenges and opportunities

John Marioni

European Bioinformatics Institute, Hinxton UK
marioni@ebi.ac.uk

Exploring the transcriptome at the single-cell level will provide key insights into biological processes ranging from early development to tumour etiology. While recent technological advances have allowed single-cell transcriptomes to be generated via next-generation sequencing, many computational challenges have to be overcome to make best use of the data generated. In this presentation I will discuss some of the technical issues that arise in single-cell transcriptomics before discussing various computational approaches for interpreting the data.

Abstracts of poster presentations

[in alphabetical order]

Computational phenotype prediction of ionizing-radiation-resistant bacteria with a multiple-instance learning model

Sabeur Aridhi
LIMOS - UBP - Clermont University, BP 10125, 63173, Clermont Ferrand, France.
LIPAH - FST - University of Tunis El Manar, Tunis, Tunisia.

Haitham Sghaier
Unit of Microbiology and Molecular Biology, National Center for Nuclear Sciences and Technologies (CNSTN), Sidi Thabet Technopark,2020 Sidi Thabet, Tunisia

Mondher Maddouri
LIPAH - FST - University of Tunis El Manar, Tunis, Tunisia
Engelbert Mephu Nguifo
LIMOS - UBP - Clermont University, BP 10125, 63173, Clermont Ferrand, France

Ionizing-radiation-resistant bacteria (IRRB) are important in biotechnology. The use of these bacteria for the treatment of radioactive wastes is determined by their surprising capacity of adaptation to radionuclides and a variety of toxic molecules. In silico methods are unavailable for the purpose of phenotypic prediction and genotype-phenotype relationship discovery. We analyzed basal DNA repair proteins of most known proteomes sequences of IRRB and ionizing-radiation-sensitive bacteria (IRSB) in order to learn a classifier that correctly predicts unseen bacteria. In this work, we formulated the problem of predicting IRRB as a multiple-instance learning (MIL) problem and we proposed a novel approach for predicting IRRB.
We used a local alignment technique to measure the similarity between protein sequences to predict ionizing-radiation-resistant bacteria. The first results are satisfactory and provide a MIL-based prediction system that predicts whether a bacterium belongs to IRRB or to IRSB. The proposed system is available online at http://home.isima.fr/irrb/.

Efficient network-guided multi-locus association mapping with graph cuts

Chloe-Agathe Azencott 1, Dominik Grimm 1, Mahito Sugiyama 1, Yoshinobu Kawahara 2 and Karsten M. Borgwardt 1;3

1 Machine Learning and Computational Biology Research Group, Max Planck Institute for Developmental Biology & Max Planck Institute for Intelligent Systems Spemannstr. 38, 72076 Tubingen, Germany
2 The Institute of Scientific and Industrial Research (ISIR) Osaka University 8-1 Mihogaoka, Ibaraki-shi, Osaka 567-0047 Japan
3 Zentrum fur Bioinformatik, Eberhard Karls Universitat Tubingen, 72076 Tubingen, Germany

As an increasing number of genome-wide association studies reveal the limitations of the attempt to explain phenotypic heritability by single genetic loci, there is a recent focus on associating complex phenotypes with sets of genetic loci. While several methods for multi-locus mapping have been proposed, it is often unclear how to relate the detected loci to the growing knowledge about gene pathways and networks. The few methods that take biological pathways or networks into account are either restricted to investigating a limited number of predetermined sets of loci, or do not scale to genome-wide settings.

We present SConES, a new efficient method to discover sets of genetic loci that are maximally associated with a phenotype, while being connected in an underlying network. Our approach is based on a minimum cut reformulation of the problem of selecting features under sparsity and connectivity constraints, which can be solved exactly and rapidly. SConES outperforms state-of-the-art competitors in terms of runtime, scales to hundreds of thousands of genetic loci and exhibits higher power in detecting causal SNPs in simulation studies than other methods. On flowering time phenotypes and genotypes from Arabidopsis thaliana, SConES detects loci that enable accurate phenotype prediction and that are supported by the literature.

Note: Paper accepted for presentation at ISMB/ECCB 2013

Using approximate Bayesian inference for gene set analysis integrating multilevel omics data

Florian Buettner [1], Steffen Sass [1], Nikola S. Mueller [1] and Fabian J. Theis [1;2]

[1]Institute of Computational Biology, Helmholtz Zentrum Munchen, Ingolstadter Landstraße 1, 85764 Neuherberg, Germany
[2]Department of Mathematics, Technische Universitat Munchen, Boltzmannstraße 3, 85747 Garching, Germany

Modern high-throughput methods allow the investigation of biological functions across multiple ``omics'' levels. Levels include mRNA and protein expression profiling as well as additional knowledge on e.g. DNA methylation and microRNA regulation. The reason for this interest in multi-omics is that actual cellular responses to different conditions are best explained mechanistically when taking all omics levels into account. To map gene products to their biological functions, public ontologies like Gene Ontology (GO) are commonly used. Many methods have been developed to identify terms in an ontology, overrepresented within a set of genes. However, these methods are not able to appropriately deal with any combination of several data types. Here, we propose a new method to analyse integrated data across multiple omics-levels to simultaneously assess their biological meaning. We developed a model-based Bayesian method for inferring interpretable term probabilities in a modular framework. Our Multilevel ONtology Analysis (MONA) algorithm performed significantly better than conventional analyses of individual levels and yields best results even for sophisticated models including mRNA fine-tuning by microRNAs.

A mixed-model approach for association studies of multiple variants to a set of correlated traits

Francesco Paolo Casale1, Barbara Rakitsch2, Christoph Lippert3, Oliver Stegle1

1 EMBL-European Bioinformatics Institute, Hinxton, Cambridge, UK
2 Max Planck Institutes Tubingen, Tubingen, Germany
3 Microsoft Research, Los Angeles, California, USA

Linear models are a key tool to investigate genotype-to-phenotype relationships as done in genome-wide association studies (GWAS). To date, large-scale studies have helped to identify hundreds of loci in the genome that are in association with disease phenotypes and other quantitative traits of interest. While GWAS have been applied to a range of different biological systems, their computational analysis remains challenging. First, while sample sizes are typically small compared to the number of individual genetic variants to be considered, true biological effects are often precluded by population structure, causing non-IID sample structure. Second, individual SNPs frequently explain little variance in isolation and hence need to be considered jointly to detect sufficient effects. Finally, individual traits are increasingly measured jointly and approaches to combine related phenotypes to increase power are thus needed.
Here, we present a random effect model to address these open needs. We jointly model groups of related phenotypes, attributing their variability to a focused region of the genome while accounting for background signals from the entire genome. In simulated settings, we show that this approach performs better than existing multi phenotype models, ignoring signals of multiple genetic variants. Finally, in the context of retrospective analysis of flowering phenotype data from A. thaliana, we show how the fitted model parameters can be interpreted to dissect the overall phenotypic variance into interpretable genetic factors.

Analysis and Interpretation of Big Imaging Mass Spectrometry Data by Clustering Mass-to-charge Images According to their Spatial Similarity

Ilya Chernyavsky1, Sergey Nikolenko1;2;3, and Theodore Alexandrov4;5;6

1 St. Petersburg Academic University, St. Petersburg, Russia
2 National Research University Higher School of Economics, St. Petersburg, Russia
3 Steklov Mathematical Institute, St. Petersburg, Russia
4 Center for Industrial Mathematics, University of Bremen, Germany
5 Steinbeis Innovation Center SCiLS, Bremen, Germany
6 Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
theodore@uni-bremen.de

Imaging mass spectrometry (IMS) has emerged in the last decade as a label-free, spatially-resolved, and multi-purpose bioanalytical technique for direct analysis of biological samples from animal tissue, plant tissue, bio- and polymer films. Analysis and interpretation of an IMS dataset is a complicated endeavor because of its large size: a dataset usually comprises thousands of spectra and tens to hundreds thousands of mass-to-charge (m/z) images. We propose a novel and easy-to-implement approach to answer the key question of unsupervised analysis of IMS data: what do all m/z-images look like? The key idea is to cluster all m/z-images based on their spatial similarity so that each cluster contains spatially-similar m/z-images. We evaluate our approach on a matrix-assisted laser desorption/ionization (MALDI) IMS dataset of a rat brain coronal section and propose a visualization method for both spatial and spectral information that allows one to quickly understand how all m/z-images look.

Machine Learning Driven Prediction of Pathogenicity

Carlus Deneke, Bernhard Y Renard

Robert Koch-Institut, Berlin, Germany
RenardB@rki.de

Introduction

In recent years, the number of sequenced and annotated organism has continuously increased and a large variety of pathogenic organisms have been collected. It has become a major goal of virologists and bacteriologists to pinpoint those sequences that ultimately induce pathogenicity. Since information of large data sets needs to be processed, methods from bioinformatics and statistical learning are in strong demand to support this process. This contribution reports on the application of machine learning for the classification of virulence-related protein sequences in bacteria.

In the following, we outline our strategy for data compilation and feature extraction as well as the methods of statistical learning applied. We conclude by summarizing our main results including a comparative study to literature data.

Data compilation

For any machine learning task, it is essential to define proper training and test sets. Manually curated lists of virulent proteins are available at several public databases. For the present purpose, we use all entries of Gammaproteobacteria in the virulence factor data base (VFDB) (1) to define the positive data set. Data for the negative set can be retrieved from various online resources such as Uniprot. Since we would like to classify proteins involved in pathogenicity in humans, we restrict the negative data set to all Gammaproteobacteria that are actually found in humans (i.e. are contained in the human microbiome) and are not annotated as virulent. Previous studies only compared proteins in entire bacterial classes or phyla regardless of their linkage to humans. The refinement undertaken here therefore assures that the learning task is performed on data relevant to humans.

Features

From these data, we generate a variety of features based on protein sequences. Concretely, we consider features like the residue and di-peptide count statistics, sequence correlations and discriminant sequence motifs that can be directly inferred from the protein sequence data. Further features are extracted from the physico-chemical properties and the predicted secondary structure. Additionally, features based on sequence comparison, i.e. the relatedness to humans and other organisms, may also increase the discriminative power. The inclusion of not directly accessible information such as secondary structure information is a major novelty in comparison to existing approaches.

Learning Strategy

We implement various supervised learning strategies and evaluate their respective advantages. Since we deal with imbalanced data (few virulent examples vs. many non-virulent sequences) and a large feature space, random forests (2) as well as support vector machines are very suitable learners. We can furthermore discern the most relevant features via feature importance which in turn aids the biological interpretation of the findings.

In many real-world situations, one faces the problem that labeled data are scarce but unlabeled data abound. In the present case of virulence classification, only relatively few proteins have high evidence to be virulent whereas the status of the majority of sequenced proteins remains in doubt. Similarly, proteins not labeled as virulent may lack this quality due to incomplete annotation. Therefore, to account for these uncertainties, we additionally experiment with a semi-supervised learning strategy.
Figure 2 illustrates the central idea: A relatively low number of sequences is initially labeled while the majority of the data is unlabeled. Based on the labeled data alone the decision boundary could only be estimated with great uncertainty. However, the inclusion of the unlabeled data via semi-supervised classification techniques such as self- and co-learning (3) allows labeling all data and ideally to a much better prediction accuracy of the test set. Semi-supervised learning is still a relatively young topic with many developments in recent years. Although it appears a natural extension in order to include a larger training set, this contribution shows the first application to virulence classification.

Results and Conclusion

The application of the supervised learning strategy yields low classification errors with sensitivity and specificity well above 80 %. Previous studies applied machine learning techniques for the classification of adhesins (4), bacterial toxins (5) and bacterial virulence factors (6). As a benchmark test, we applied our learning strategy to the same data sets as in these publications and we succeeded to yield competitive results in all cases.

Our results underline the importance of the novel aspects introduced in this contribution: Firstly, we addressed the compilation of adequate data more precisely as previously done. Secondly, we devised an array of features related to the secondary structure of proteins which aim to model the protein's function more precisely. Furthermore, we extended the methodology of the learning process by employing a semi-supervised strategy. Thus, the approach presented here describes a powerful method to discriminate previously unknown protein sequences in respect of their potential to affect the virulence of an organism. Our method is effective and can be run on any standard desktop computer. This is advantageous as this allows an implementation as a fast and reliable diagnostic of newly discovered potential pathogens.

Inferring mechanisms of combinatorial regulation from thousands of dose-response curves of designed promoters

David van Dijk, Eilon Sharon & Eran Segal
Weizmann Institute of Science

Gene expression is a function of promoter sequence and the activity of regulating transcription factors (TFs). However, it is not well understood how regulation, through promoter sequence changes, depends on TF activity. Here we measure the dose-response curve of 6500 designed promoters in a serial amino acid starvation in yeast. The promoters are target of various amino acid induced TFs, such as Gcn4 and Leu3. To investigate combinatorial regulation, we vary the position, number and affinity of the binding sites as well as their accessibility, by modulating the nucleosome occupancy (polyT content) of the promoters. We find that adding binding sites increases the dynamic rage of expression and adding nucleosome-disfavoring sequences mostly shifts expression without changing the dynamic range. In addition we observe a synergistic effect of two binding sites, where gene expression is higher than the sum of the expression of the single sites, and an anti-synergistic effect where adding an activator site reduces expression.

From sequence to enzyme mechanism

Luna De Ferrari and John B O Mitchell

Biomedical Sciences Research Complex and EaStCHEM School of Chemistry, Purdie Building, University of St Andrews, North Haugh, St Andrews,
Scotland KY16 9ST, UK
Email: luna.deferrari@st-andrews.ac.uk

Background:
Predicting enzyme function at the level of chemical mechanism provides a finer granularity of annotation than traditional Enzyme Commission (EC) classes. Predicting not only whether a putative enzyme in a newly sequenced organism has the potential to perform a certain reaction, but how the reaction is performed, using which cofactors and with susceptibility to which drugs or inhibitors, has important consequences for drug and enzyme design. Work such as SABER (Nosrati et al. 2012) predicts enzyme catalytic activity based on 3D protein features. However, using 3D structural attributes limits the prediction of mechanism to proteins already having either a solved structure or a close relative suitable for homology modelling.

Results:
In this study, we evaluate whether InterPro and Catalytic Site Atlas sequence signatures provide enough information for bulk prediction of enzyme mechanism. The method can predict at 89% accuracy (92% micro-averaged precision, 99.7% macro-averaged recall) the MACiE mechanism definitions of 288 proteins available in the EzCatDb, MACiE and SFLD databases using an off-the-shelf multi-label K-Nearest Neighbours algorithm.

Conclusion:
We find that InterPro signatures are critical for accurate prediction of enzyme mechanism, showing robust accuracy even when a reduced subset of signatures containing enzyme names are used. We also find that incorporating Catalytic Site Atlas attributes results in little additional accuracy. The software code (ml2db), data and results are available online at http://sourceforge.net/projects/ml2db/.

Gene Regulatory Network Inference using ensembles of Local Multiple Kernel Models

Arnaud Fouchet
email:afouchet@ibisc.fr

Jean-Marc Delosme
email:jean-marc.delosme@ibisc.fr

Florence d'Alche-Buc
email:florence.dalche@inria.fr

INRIA-Saclay, AMIB / TAO, LRI umr 6623 CNRS, Universite Paris Sud
91150 Orsay, France
IBISC, EA 4526, Universite d'Evry-Val d'Essonne
91037 cedex Evry, France

Reconstructing gene regulatory network from high-throughput data has many potential applications, from understanding a biological organism to identifying potential drug targets. It is also a notoriously difficult problem, tackled by many scientists with various methods
In this paper, we formulate GRN inference as a sparse regression problem. We decompose the prediction of a p-genes system in p different regression problems. For each gene (target gene), we train a kernel-based regression with feature selection, predicting the expression pattern of the target gene using all the other genes (input genes). The regression will give the importance of each input gene in the prediction of the target gene. We take this importance as an indication of a putative regulatory link. Putative links are the aggregated over all genes to provide a ranking of interactions, from which we infer the GRN.
Furthermore, biological data are heterogeneous. The method we propose can learn from both steady-state and time-series data, using an ensemble method that can be applied to other regression model.
Finally, we compare our method, called LocKING, to state-of-the-art methods on real and realistic datasets, which are widely spread in the GRN inference community. We show that our method is competitive against individual methods. Nevertheless, best results are obtained by integrating multiple methods. We show that using LocKING among other methods significantly enhances the accuracy of the network inferred.

MINT: Mutual Information based Transductive Feature Selection for Genetic Trait Prediction

Dan He1, Irina Rish1, David Haws1, Simon Teyssedre2, Zivan Karaman2, Laxmi Parida1

1 IBM T.J. Watson Research, Yorktown Heights, NY, USA
{dhe, rish, dhaws, parida}@us.ibm.com
2 Limagrain Europe, Chappes Research Center, CS 3911, 63720 Chappes, France
fsimon.teyssedre, zivan.karamang@limagrain.com

Whole genome prediction of complex phenotypic traits using high-density genotyping arrays has recently attracted a lot of attention, as it is relevant for the fields of plant and animal breeding and genetic epidemiology. Given a set of biallelic molecular markers, such as SNPs, with genotype values encoded as {0, 1, 2} on a collection of plant, animal or human samples, the goal is to predict the values of certain traits, usually highly polygenic and quantitative, by modeling simultaneously all marker effects, unlike the traditional GWAS. As the number of genotypes is generally much bigger than the number of samples, the predictive models suffer from the "curse of dimensionality". In this work, we proposed a transductive feature selection method MINT based on information theory. MINT applies the MRMR (Max-Relevance and Min-Redundancy) criterion and integrates the test data in a natural way in the feature selection process. A dynamic programming algorithm is developed to speed up the selection process. Our experiments on both simulated and real data show that MINT generally achieves similar or better results than the state-of-the-art method mRMR does which is also based on the MRMR criterion and relies on training data only. To our knowledge, this is the first transductive feature selection method based on the MRMR criterion.

Nested sampling for Bayesian model selection and parameter inference

Rob Johnson, Paul Kirk, Michael P. H. Stumpf

Center for Bioinformatics, Imperial College London, London, UK / Institute of Mathematical Sciences, Imperial College London, London, UK

A common practice in systems biology is to construct models that try to capture or explain some biological behaviour of a system. When many such models are proposed for a single system, it is pertinent to ask if one model is `better' than another - better able to explain data representative of the behaviour we are trying to model, a task known as model selection.

In the Bayesian framework, we have a formulation for independently quantifying the support afforded to a model by some data. This metric we call the evidence. It is ratios of evidences, or Bayes factors, that indicate relative support for one model over another.

Nested sampling is a method for estimating a model's evidence (Skilling, 2006). Through iterative sampling of the prior subject to nested likelihood constraints, the evidence is accumulated and the posterior parameter distribution is sampled.

We present a C-based tool for computational biologists that performs nested sampling. We test our package on both real and synthetic signalling network models (Swameye, 2003; Vyshemirsky, 2008).

Network inference via automated ODE model construction

Paul Kirk and Michael P. H. Stumpf

Center for Bioinformatics, Imperial College London, London, UK / Institute of Mathematical Sciences, Imperial College London, London, UK

The construction of models that describe the dependencies between interacting molecular species remains at the heart of systems biology. Here we develop a method for performing network inference by automatically searching through possible ODE models of biomolecular systems. Combinatorial complexity means that methods that adopt an exhaustive search of all network topologies will be infeasible, so we propose an approach that decouples the system of ODEs and hence allows us to consider each node of the network individually and in parallel. We are therefore able to perform network inference by determining the parents of each node in turn. Adopting the same approach as is used by gradient matching techniques for ODE parameter estimation, we start by first obtaining data-driven estimates (using splines or Gaussian process regression) of each of the individual state variables as a function of time. We assume that each node may have a maximum of M parents (where M << p, the total number of nodes), and then -- for each node -- we exhaustively consider all possible parent combinations. Crucially, the use of data-driven estimates of the state variables means that we do not need to consider the full system of ODEs each time we propose a new set of parents, which means that computation is feasible. There are many potential applications of this approach; we will demonstrate several. Of particular current relevance, we will use the method to identify points of cross-talk between canonical signalling pathways.

Supervised and unsupervised biological network inference from multiple 'omic data

Jana Kludas, Fitsum Tamene, Juho Rousu

Helsinki Institute for Information Technology
Department of Information and Computer Science
Aalto University, Espoo, Finland
[jana.kludas,fitsum.tamene,juho.rousu]@aalto.fi

Our goal is to investigate and ultimately improve biological network reconstruction based on integrating proteomic and genomic data. First, this work investigates the descriptive power of different features extracted from the protein sequences and genes for predicting biological networks of protein-protein interactions and metabolic networks. Three fundamentally different methods for graph inference have been implemented: (I) classification and integration of local interaction models, (II) classification of a global interaction model and (III) an unsupervised method based on estimating the inverse covariance matrix. The results show that for PPI and metabolic network prediction different data sources and different learning strategies are effective. For predicting the metabolic network the integration of proteomic and genomic data sources performs best. On the other hand, protein-protein interactions are best predicted by classification of sequence alignment scores.

Multivariate Markov Models for Precise Identification of MicroRNA Binding Sites

William H. Majoros, Parawee Lekprasert, Neelanjan Mukherjee, Rebecca L. Skalsky, David L. Corcoran, Bryan R. Cullen, Uwe Ohler

Institute for Genome Sciences & Policy, Duke University, Durham NC, USA
Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham NC, USA
Berlin Institute for Medical Systems Biology, Max Delbruck Center, Berlin Germany
uwe.ohler@mdc-berlin.de

High-throughput sequencing has dramatically improved our ability to interrogate various aspects of cellular state, including binding events of various types. Within the context of gene regulation, binding of microRNAs to gene transcripts has been shown to play an enormously important role in modulating expression levels, with mounting evidence that the dysregulation of microRNA binding can result in a range of human diseases. Up until now, assays for microRNA binding in vivo have not directly identified which microRNA is involved in any given binding event, and they do not yet operate at single-nucleotide resolution. We show that via the joint modeling of (Argonaute) protein-RNA crosslinking, microRNA sequence binding preferences, and evolutionary conservation, microRNA binding sites can be precisely located and the identity of the involved microRNA family can be correctly deduced, with very high accuracy. We perform this joint modeling via a 47-state multivariate Markov model having both discrete and continuous emissions. We have found this class of models to be highly flexible, and we show how our microRNA binding model was easily modified to additionally investigate the prevalence of recently-proposed alternate binding modes. Very preliminary results suggest that similar models can be useful for additional modeling tasks for RNA-binding proteins, such as localized motif discovery and investigation of combinatorial binding. In addition, we believe this class of models will prove to have utility for precise identification of regulatory binding events from other data sources, such as DNase, MNase, and ChIP data for DNA-binding factors.

Predicting Unique E2f1 and E2f4 Targets Using the DNA Binding Profiles of Transcription Factors and Modified Histones

[1,&]Dinesh Manandhar, [1,&]Adegoke Ojewole, [1,2,]*Raluca Gordân

[1]Institute for Genome Sciences and Policy,
[2]Departments of Biostatistics and Bioinformatics, Computer Science, and Molecular Genetics and Microbiology, Duke University, Durham, NC 27708, USA
[&]These authors contributed equally to this work
*Corresponding author. Email: raluca.gordan@duke.edu

Members of the E2f family of transcription factors (TFs) mediate a wide range of functions in eukaryotic cells, from DNA synthesis and repair to cell cycle regulation and apoptosis. E2f1 and E2f4 are of particular interest because, despite their shared binding specificity, each protein has unique targets and unique regulatory functions in the cell: E2f1 acts mainly as a transcriptional activator and plays important roles during apoptosis and cell-cycle progression from G1 to S phase, while E2f4 acts mainly as a transcriptional repressor. The intrinsic DNA binding properties of E2F1 and E2F4 cannot explain how the two TFs identify their unique targets in the genome. In this study, we use the in vivo DNA binding profiles of putative co-factors and histone modifications to predict unique E2f1 and E2f4 genomic targets, using Support Vector Machine (SVM) and Random Forest (RF) classifiers. Using chromatin immunoprecipitation (ChIP-seq) data for 52 TFs, our approach predicts unique E2f1 and E2f4 binding sites with ~84% accuracy. Using ChIP-seq data for 11 modified histones we obtain a classification accuracy of ~69%. Combining both TF and histone modification ChIP-seq data, our accuracy increases to ~86%. Importantly, some of the features that contribute most to the classification accuracy have already been shown to interact with E2f1 and/or E2f4. Thus, our results suggest that the genomic environment in the neighborhood of putative E2f target sites may play an important role in the recruitment of specific E2F family members.

Interfacing cellular networks of *S. cerevisiae* and *E. coli*

Ricardo de Matos Simoes[1], Matthias Dehmer[2], Frank Emmert-Streib[1]

[1] Computational Biology and Machine Learning Laboratory
Center for Cancer Research and Cell Biology
School of Medicine, Dentistry and Biomedical Sciences
Faculty of Medicine, Health and Life Sciences
Queen's University Belfast
97 Lisburn Road, Belfast, UK
[2] Institute for Bioinformatics and Translational Research
UMIT, Hall in Tyrol, Austria

The concerted interactions on all cellular levels between genes and their gene products within a cell are governed by various types of gene networks. However, the relation and the biological overlap among experimental and inferential gene networks, e.g., between the experimental transcriptional regulatory network (TRN), the experimental protein interaction network (PPN) and the inferential gene regulatory network (GRN) that is inferred from large-scale transcriptomic data is largely unknown. In order to develop integrative methods that consider multiple levels of the gene network it is crucial to investigate global structural similarities and functional roles of the network interfaces that are defined by the subnetwork of shared interactions. We provide in this study an in-depth analysis of the structural, functional and chromosomal relationship between a protein-protein network, a transcriptional regulatory network and an inferred gene regulatory network for \textit{S. cerevisiae} and E. coli (Figure 1 A). We infer gene regulatory networks by using the BC3NET (de Matos Simoes 2012) from large-scale gene expression compendium for S. cerevisiae and E. coli (Faith 2008). The protein networks consider physical protein interactions from various large-scale protein interaction databases for S. cerevisiae (Wu 2009) and for E.coli (Xenarios 2002, Aranda 2010, Licata 2012, Goll 2008). For the transcription regulatory interactions that are used in our study are based on high-throughput experiments (e.g. ChIP) for S. cerevisiae (Balaji 2006) and for E. coli (Gama-Castro 2011). We study global and local aspects of these networks and their biological information overlap by comparing, e.g., the functional co-occurrence of Gene Ontology terms by exploiting the available interaction structure among the genes. The functional analysis are performed based on classical and a variety of global and local network based Gene Ontology analysis startegies.

Discovery of multi-variant effects in complex diseases: an opportunity for machine learning

Michael A. Mooney[1,*], Bipolar Genome Study, Psychiatric Genomics Consortium, Shannon K. McWeeney[1]

[1] Department of Medical Informatics & Clinical Epidemiology, Division of Bioinformatics & Computational Biology, Oregon Health & Science University
[*] mooneymi@ohsu.edu

The genome-wide association study (GWAS) is an important method for identifying genes involved in complex diseases. However, the genetic associations discovered so far account for only a small part of the genetic component of the diseases studied. In order to create a more complete understanding of genetic risk factors for complex traits, polygenic (multi-variant) analyses are becoming an important part of genomic studies.
Given the large amount of genetic heterogeneity among patients with complex diseases it is important to take into account the effects of multiple (possibly many) genetic variants when searching for genetic signatures of disease risk. Likewise, it is important to examine the possibility that variants will have different effects, for instance main effects versus interactive (conditional) effects, on the trait of interest. The large number of variants now being measured makes an exhaustive search of all possible variant combinations impossible. Machine learning techniques, most likely in combination with various forms of expert knowledge (functional predictions, gene-gene interactions, etc.), will allow researchers to investigate more complex and more biologically informative genetic signatures of risk for complex diseases.
Here we describe the application of a network-guided probabilistic feature-selection algorithm to polygenic analyses in two complex disease data sets. The goals of this study were: 1) to show the diversity of analyses (or models) that can be used to extract information from the wealth of genomic data currently available; and 2) to provide evidence for the necessity, and the feasibility, of using machine learning approaches to mine this data and to generate meaningful biological hypotheses about the diseases under study.

Efficient Determination of the conditonal dependence of protein subcellular localization by Active Learning

Armaghan Naik, Joshua Kangas, Christopher Langmead and Robert F. Murphy

Lane Center for Computational Biology, Joint Carnegie Mellon University-University of Pittsburgh PhD. Program in Computational Biology, Depts. of Biol. Sciences, Biomed. Engineering & Machine Learning
murphy@cmu.edu

High throughput and high content screening involve determination of the effect of many chemical compounds on a given cellular target. As currently practiced, a full set of measurements for all compounds for each new target is typically made with little use of information from previous screens. To efficiently study many targets and their dependences on small molecules, a means is needed for determining and exploiting similarities in the effects of compounds and/or behavior of targets such that a smaller number of measurements can be made while preserving high accuracy. Here, we describe probabilistic models that can be used to predict results for unmeasured combinations, and active learning algorithms for efficiently selecting future informative batches of experiments. Through extensive simulated experiments we show that our approaches can produce powerful predictive models and learn them significantly faster than can be done by random choice. We further characterized our method's performance for learning the dependence of subcellular location of proteins in a collection of 96 NIH-3T3 clones, each endogenously expressing a GFP tagged protein with a library of 96 small molecule compounds. Our method achieved a 92% accuracy having only sampled 28% of the experiment space ($P<10^{-84}$)

QUANTITATIVE GLOBAL ANALYSIS OF ENZYME REACTION MECHANISMS

Neetika Nath, Lazaros Mavridis & John B. O. Mitchell

Biomedical Sciences Research Complex and EastCHEM School of Chemistry, Purdie Building,
University of St Andrews, North Haugh, St Andrews, Scotland KY16 9ST, UK.

Motivation: Global study of enzyme reaction mechanisms may provide important insights for better understanding of the diversity of chemical reactions of enzymes. In this study, we describe how the chemical mechanisms of enzyme reactions cluster in a space defined by chemoinformatics descriptors, using unsupervised global analysis.

Results: We have performed a clustering analysis of enzyme reactions described first by overall similarity (OS) descriptors, and second by mechanistic similarity (MS) descriptors. We have designed a clustering algorithm, PFClust, that groups enzyme overall reactions by OS into 34 clusters, and enzyme reaction mechanisms by MS into 21 clusters.

Conclusion: We find that the MS descriptors cluster the data into significantly fewer clusters than OS descriptors. Similar mechanisms more frequently cluster together than do similar overall reactions. This means that different functions tend to share similar mechanisms.

Statistical Analysis of the Evolution of Networks

Gunter Neumann

Friedrich-Alexander-Universitat,Department Mathematik
Cauerstr. 11, 91058 Erlangen, Germany
Email: gunter.neumann@fau.de

We want to link different levels of biological networks in order to explain evolutionary aspects. Sequence comparison, next to energetic metabolic network analysis and expression profiles are compared to derive interactions and interdependencies. Environmental aspects are considered under the optimization of metabolic yield and optimality of network structure while also including the interaction of different species. Organisms are classified according to their type of interactions with the environment.

We use theoretical observables to classify and characterize interaction dynamics by Network analysis. Therefore sub networks/modules accessible to the continuous dynamical systems analysis are characterized by their structural stability. A combination of biological arguments and systems dynamical criterions are then used to define a biological network as evolutionary stable if subsystems obey evolutionary stability. A thermodynamical description of the metabolic network governed by the regulatory network is considered under the aspect of minimization of energetic effort. However different strategy adaptations from organisms lead to structurally different organizations of networks.

For example sequence analysis is included into the description of networks. Aspects such as flexibility and node degree in the interaction network are studied. The connection to the function is then analysed. A trade off between structure and flexibility is created that should represent the relation between speed of evolution and accuracy of the definition of the function of the sequence under consideration. The “age” of an organism is defined as the time where an organism doesn’t change its DNA or function more than a threshold value. Also the age of an organism is compared to its stability. Species with different ages are then compared by their network structure and consistency of the corresponding transcribed proteins/RNA.

Furthermore we want to partition the network into different classes. The resulting network is multi modal. The definition of the membership of a node (for example a metabolite, RNA, protein) is defined by the role it plays in the network. This is important to compare different networks since there often the same function is adopted by different substrates.

Another reasonable interface between theory and biology can be obtained by comparing networks from pathological organisms with evolutionary stable considered structures to obtain a system theoretic explanation for the occurrence of a genetic disease.

Automated annotation of gene expression image sequences via nonparametric factor analysis and conditional random fields

Iulian Pruteanu-Malinici, William Majoros and Uwe Ohler

Institute for Genome Sciences & Policy, Duke University, Durham NC, USA
Berlin Institute for Medical Systems Biology, Max Delbruck Center, Berlin Germany
uwe.ohler@mdc-berlin.de

Computational approaches for the annotation of phenotypes from image data have shown promising results across many applications, and provide rich and valuable information for studying gene function and interactions. While data are often available both at high spatial resolution and across multiple time points, phenotypes are frequently annotated independently, for individual time points only. In particular, for the analysis of developmental gene expression patterns, it is biologically sensible when images across multiple time points are jointly accounted for, such that spatial and temporal dependencies are captured simultaneously.

We describe a discriminative, undirected graphical model to label gene-expression time-series image data, with an efficient training and decoding method based on the junction tree algorithm. The approach is based on an effective feature selection technique, consisting of a nonparametric sparse Bayesian factor analysis model. The result is a flexible framework, which can handle large-scale data with noisy, incomplete samples, i.e. it can tolerate data missing from individual time points.

Using the annotation of gene expression patterns across stages of Drosophila embryonic development as an example, we demonstrate that our method achieves superior accuracy, gained by jointly annotating phenotype sequences, when compared to previous models that annotate each stage in isolation. The experimental results on missing data indicate that our joint learning method successfully annotates genes for which no expression data are available for one or more stages.

Note: Paper accepted for presentation at ISMB/ECCB 2013

A PYTHON LIBARY FOR PROBABILISTIC GRAPHICAL MODELS, WITH APPLICATIONS TO BIOINFORMATICS

EMANUELE RAINERI, CNAG Spain
emanuele.raineri@gmail.com

Many computations in bioinformatics can be seen as maximizing, or calculating marginals of, or conditioning on probability distributions (e.g., this happens when using an HMM to align sequences).

Furthermore, many concepts in biology have a natural representation as graphs of interacting molecules (signalling networks, flow of genes in pedigrees, etc).

This poster presents a library which allows one to define probabilistic graphical models and evaluate them, together with examples of applications to computational biology.

The user can specify a factor graph or a belief network (possibly with loops) and employ exact message passing algorithms to determine the maximum and the marginals of the corresponding probability distribution (the belief network is internally converted into a junction tree for efficiency).

I will illustrate how this library can be used with three examples : inferring the DNA sequence of a family member having sequenced other members of the pedigree distinguishing de novo mutations from sequencing errors, and probabilistic inference on protein networks.

Supervised Learning Methods to Infer Metabolic Network using Sequence and Non-sequence Kernels

Abiel Roche-Lima, Michael Domaratzki, Brian Fristensky

Bio Information Technologies Lab., University of Manitoba, Winnipeg, MB R3T 2N2, Canada
email contact: aroche@cs.umanitoba.ca

Metabolic networks allow the modelling of molecular systems to understand the underlying biological mechanisms in a cell, represented by the set of metabolic pathways. In living cells, a series of chemical reactions are carried out in a stepwise manner by different proteins called enzymes. However, many biochemical pathways remain incompletely characterized, and in some pathways, not all enzyme components have been identified.

Kernel methods have proven useful in many difficult problems such as document classifications and bioinformatics. A kernel is a measure of similarity that satisfies the additional condition of being a dot product in some feature space.

In this work, we apply kernel methods to infer metabolic networks, considering non-sequence and sequence kernels. Non-sequence kernels describe feature spaces for information that is not directly related to sequence data, e.g., protein localization and phylogenetic data. Sequence kernels include similarity as a set of sequence models, and each component of the feature space representation measures the extent to which a given sequence fits the model.

Recently, supervised network inference methods for predicting biological networks have been developed in the framework of kernel methods. We use two of these methods in our research: support vector machines (SVMs) and Kernel regressions. SVMs are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Kernel regression methods are based on the supervised graph inference framework to infer metabolic networks with metric learning.

In this work, we aim to compare sequence and non-sequence kernels, using SVM and kernel regression methods to infer metabolic networks. We develop several experiments to accomplish it.

Dimensionality reduction for high-content screening data

Lee Zamparo_1,2 and Zhaolei Zhang†2

1Department of Computer Science, University of Toronto
2Donnelly Centre for Cellular and Biomolecular Research, University of Toronto
Contact: zamparo@cs.toronto.edu

Motivation: High content screening experiments offer a genome wide exploration of genetic interaction or function, by observing isogenic populations, each subject to distinct genetic perturbations, with high-throughput fluorescence microscopy. Two important tasks in high content screening, often called high content imaging, are to identify cells bearing interesting phenotypes, and to estimate the number of distinct phenotypes present in a given screen. A common approach to these problems is to transform the images from pixels into a space of features, to perform dimensionality reduction on that space, and subsequently to build a classification or clustering model on the lower dimensional data. The success of this approach is dependent on a faithful preservation of relationships between features in the reduced space, and current popular approaches such as PCA perform poorly on data with non-linear relationships between covariates.
Results: We propose, construct and train a deep learning auto-encoder to learn a function for dimensionality reduction of data derived from cell images. This has several advantages over competing methods, including an increased capability to model the relationships in the data, no costly out-of-sample problems, the ability to fine-tune the model with any available labeled data, and the ability to scale to very large data sets. Though currently a work in progress, we establish that standard methods are ill-suited to our problem. We will compare our method with each of PCA, Local Linear Embedding and ISOMAP on a DNA-damage foci screen in Saccromyces Cerevisiae.
Availability: Code will be made available after publication via github.