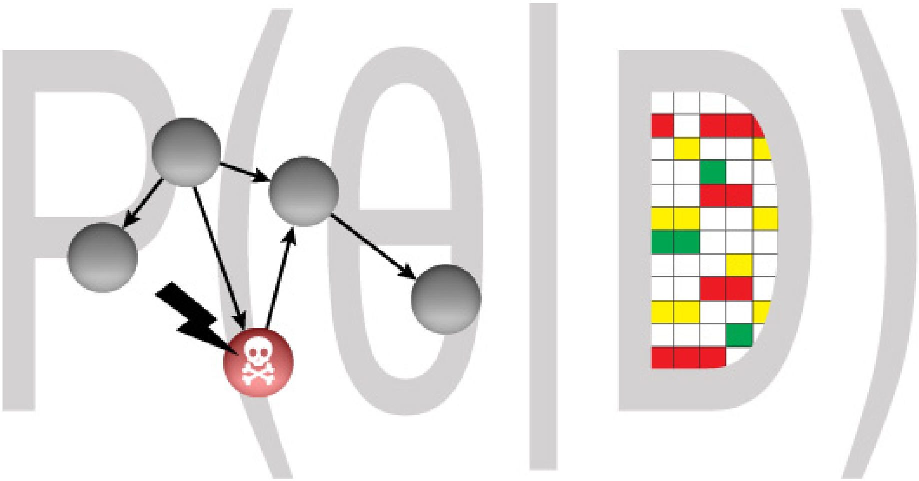
Congressi
Stefano Franscini
Swiss Federal Institute
of Technology Zurich

Statistical Learning of Biological Systems from Perturbations

Ascona, May 31 — June 5 2015



Organised by Niko Beerenwinkel, Peter Bühlman, Wolfgang Huber and Darlene Goldstein

Advances in biotechnology have made genome-scale measurements routine, including most recent techniques for perturbing individual genes in a targeted manner. These interventional data hold the promise to infer biological networks and to move forward systems biological approaches significantly. A major challenge now is to use the vast amount of data generated from these technologies and to devise appropriate statistical models and computational inference methods. Unlike observational data, interventional data can reveal causal relationships among genes or other biomolecular entities. As such, the statistical analysis and computational integration of perturbation data is an important step towards large-scale biological system identification with abundant applications in biology and medicine.

This workshop will (i) explore recent advances and open problems in statistical learning, data integration, and causal inference of biological systems; (ii) present biomedical applications to recent genome-wide perturbation data, such as RNA interference data, obtained, for example, from cancer cells or cells infected by pathogens; and (iii) facilitate meaningful interaction between biomedical and quantitative researchers.

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Roche

NOVARTIS

SIB
Swiss Institute of
Bioinformatics

Congressi
Stefano Franscini
Swiss Federal Institute
of Technology Zurich

Venue

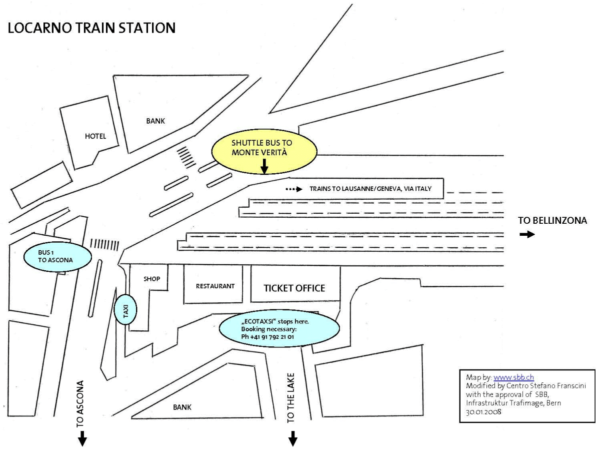
Monte Verità
Via Collina 84
CH-6612 Ascona
tel. +41 91 785 40 40

About the Congressi Stefano Franscini (CSF)

The Congressi Stefano Franscini (CSF) is the international conference centre of the Swiss Federal Institute of Technology (ETH) in Zurich, situated in the south of Switzerland (Canton Ticino) at Monte Verità. It has been named after the Federal Councillor Stefano Franscini, a native of Ticino who, in 1854, played an important part in establishing the first Federal Institute of Technology in Switzerland, ETH Zurich. Every year, the centre hosts 20 - 25 conferences organised by professors working at Swiss universities and concerning all disciplines (sciences and humanities) taught at academic level. The centre is also open to the local population with a regular program of public events (lectures, concerts, films, etc.) organised in the context of its international conferences and/or Monte Verità's cultural programme.

Shuttle service from Locarno Station

A free 13-seater shuttle bus to Monte Verità will leave from Locarno railway station Sunday May 31 at the following times: **14.05**; **14.45**; **15.25**; **16.05**; **16.45**; **17.25**. The meeting point is on the right side of the train platforms in Locarno (see image).


LOCARNO TRAIN STATION
HOTEL
BANK
SHUTTLE BUS TO MONTE VERITÀ
TRAINS TO LAUSANNE/GENEVA, VIA ITALY
TO BELLINZONA
BUS 1 TO ASCONA
SHOP
RESTAURANT
TICKET OFFICE
TAXI
„ECOTAXSI" stops here. Booking necessary: Ph +41 91 792 21 01
BANK
TO ASCONA
TO THE LAKE

Map by: www.sbb.ch
Modified by Centro Stefano Franscini
with the approval of SBB,
Infrastruktur Trafimage, Bern
30.01.2008

Keynote lectures

Genetic networks: general properties and complex phenotypes

Brenda Andrews

The Donnelly Centre, University of Toronto, Toronto, Ontario Canada

Determining how combinations of genetic variants or perturbations manifest themselves, particularly in the context of human disease, is a formidable challenge. To define the general principles of genetic networks, our group developed a unique functional genomics platform called "synthetic genetic array" (SGA) analysis that automates yeast genetics and enables the systematic construction of double and triple mutants. One of our major goals has been to use a simple phenotypic readout of cell growth rate – colony size – to produce the first complete genetic interaction map for any cell, and to empirically delineate the properties of genetic networks. Application of our automated pipeline has enabled systematic analysis of the majority of all possible 18 million yeast gene pairs. The resultant network consists of 560,000 negative and positive genetic interactions, spanning 93% of all yeast genes. Analysis of the network has revealed: [1] a central role for and unique properties of essential genes, which we consider analogous to disease-associated genes in humans; [2] hubs and pleotropic genes on the network which show a clear association with several fundamental physiological and evolutionary properties that are predictive of genetic interactions in other organisms; [3] functional modules that we use to predict and test conservation of interactions in other systems. We have also developed three powerful pipelines which combine SGA and automated microscopy for systematic and quantitative cell biological screens or phenomics. One pipeline uses SGA to introduce fluorescent markers of key cellular compartments, along with sensitizing mutations, into yeast mutant collections. We then perform live cell imaging on the mutant arrays using HTP confocal microscopy to quantitatively assess the abundance and localization of our fluorescent reporters, providing cell biological readouts of specific pathways and cellular structures in response to thousands of genetic perturbations.

Pooled shRNA screening: from large scale perturbation to single genes as hits

Roderick Beijersbergen

The Netherlands Cancer Institute

With the development of large collections of gene expression perturbing reagents such as RNAi based shRNA collections and CRIPR-based sgRNA collections, it has become possible to perform large scale functional genomic screens in pooled formats. Combined with the ability to identify each gene-modifying reagent in individual cells, one can compare the relative abundance between differnt population of cells selected for a phenotype of interest. We have explored this technology for the identification of novel drug targets in signaling pathways, for genes that enhance drug response and those that can induce cellular resistance to selected therapies. The characteristics of the reagents and the large complexity of these screen requires dedicated analysis methods to identify genes as true positive hits. I will discuss the pooled shRNA screening technology and the associated analysis methods in the context of screens aimed at the identification genotype specific lethal genes in triple negative breast cancer.

Mapping genetic interactions across multiple phenotypes and signalingstates

Michael Boutros

German Cancer Research Center (DKFZ) and Heidelberg University

Gene-gene interactions shape complex phenotypes and modify the effects of mutations during development and disease. The effects of gene-gene interactions on single phenotypes have been used to aggregate genes into functional modules. However, it has not been possible to derive directional epistatic relationships between genes that constitute regulatory networks. We used combinatorial RNA interference and phenotyping by automated imaging to generate a large genetic interaction map for 21 different phenotypic features of Drosophila cells. We devised a method that combines data on multiple phenotypes to reveal directional relationships, and report a dense regulatory network. This network could reconstruct the sequence of protein activities in mitosis, and revealed that the Ras pathway interacts with chromatin-remodelling proteins, which we show is conserved in human cells. I will also present data from genetic interaction analysis across multiple signaling states.

Morphological profiling for targeting diseases and characterizing compounds

Anne Carpenter

Broad Institute of Harvard and MIT

Microscopy images contain rich information about the state of cells, tissues, and organisms. Our laboratory is extracting patterns of morphological perturbations (”signatures”) from images in order to identify similarities and differences between various chemical or genetic treatments, with the ultimate goal to identify the causes and potential cures of disease. Using model systems that are more and more physiologically relevant, yet still compatible with automated instrumentation, we are developing assays and accompanying algorithms to extract multiparametric morphological fingerprints of cell populations.
Our goals in these profiling experiments include identifying drug mechanisms of action, the impact of disease-related alleles, performance-diverse chemical libraries, mechanisms of liver toxicity, and diagnostics for bipolar disorder and schizophrenia. We hope to make perturbations in cell morphology as computable as other large-scale functional genomics data.
The image analysis algorithms and data mining approaches we develop are freely available through the biologist-friendly open-source software, CellProfiler (www.cellprofiler.org), for both small- and large-scale experiments.

Inferring directional genetic interactions from combinatorial, multi-parametric data

Bernd Fischer

German Cancer Research Center (DKFZ)

Genes display epistatic (genetic) interactions, whereby the presence of one genetic variant can mask, alleviate or amplify the phenotypic effect of other variants. Such a directional relationship is present, for instance, if one gene product positively or negatively regulates the activity of the other, if its function temporally precedes that of the other, or if its function is a necessary requirement for the action of the other. Large-scale synthetic genetic interaction screens have been performed and have been predictive for functional relationships between genes in yeast, E. coli, C. elegans and metazoan cells. To date, all large-scale genetic interaction studies have been designed to detect gene-gene interactions based on the definition of an interaction as a departure from the combination of the genes' individual effects. This statistical definition provides limited information on the nature of the interaction, in particular, it does not reveal whether there is a directional relationship between the genes. Here, we present a method that combines genetic interactions on multiple phenotypes to reveal directional relationships, and report a dense regulatory network covering 1367 genes. It reveals the directional, temporal and logical relationships between genes and allows us to dissect regulatory networks using high-throughput intervention experimentation. Furthermore, the network could reconstruct the sequence of protein activities in mitosis, and revealed that the Ras pathway interacts with the SWI/SNF chromatin-remodelling complex, which we show is conserved in human cancer cells. Our work presents a powerful approach for reconstructing directional regulatory networks, and provides a resource for the interpretation of functional consequences of genomic alterations in disease.

Statistical challenges in the study of stability in the human microbiome

Susan Holmes

Stanford University

The human microbiome is a complex assembly of bacteria that are sensitive to many perturbations. We have developed specific tools for studying the vaginal, intestinal and oral microbiomes under different perturbations (pregnancy, hypo-salivation inducing medications and antibiotics are some examples). A suite of statistical tools written in R based on a Bioconductor package (phyloseq) allows for easy normalization, visualization and statistical testing of the longitudinal multi-table data composed of 16sRNA reads com-

bined with clinical data, transcriptomic and metabolomic profiles. Challenges we have had to address include information leaks, the heterogeneity of the data, multiplicity of choices during the analyses and validation of results.
This contains joint work with Joey McMurdie, Ben Callahan, Julia Fukuyama, Kris Sankaran and David Relman's Lab members from Stanford.

Advances in causal inference with applications to systems biology

Marloes Maathuis

ETH Zurich

I will present some recent developments in causal inference from observational data, with applications to systems biology. In the first part of the talk, I will discuss two new methods to estimate the effect of multiple simultaneous interventions (e.g., multiple gene knockouts) from observational data, under the assumption that the data come from an unknown Gaussian linear structural equation model. In the second part of the talk, I will present recent results on the estimation of causal effects via covariate adjustment. In particular, I will present a unified and complete graphical criterion to identify valid adjustment sets for directed acyclic graphs with/without hidden variables and their Markov equivalence classes.

Reconstructing regulatory networks by combining perturbation screens and steady state expression profiles

George Michailidis

University of Michigan

Reconstructing transcriptional regulatory networks is an important task in functional genomics. Different approaches have been proposed in the literature, some based on perturbation data obtained from gene knockout/knockdown experiments, while others on observational data of the underlying system in steady state. On many instances, directed acyclic graphs (DAGs) capture the structure of the underlying regulatory network. Hence, the reconstruction task corresponds to learning the DAG structure, a problem that has received a lot of attention in the literature. In this talk, we discuss this problem when prior information is available. We start by defining incomplete partial orders as a formal way for describing path-based restrictions on DAGs. Incomplete partial orders generalize the important special case of a known linear order. We then develop a stochastic optimization algorithm for obtaining estimates using BIC over the restricted space and explore its performance in synthetic and real data. Further, to obtain the prior information on the partial orders we present a framework for active learning based on an improvement by sequentially choosing the perturbation experiments. We then turn to techniques for addressing the computational challenge of estimating the proposed improvement function, taken to be the entropy of certain posterior marginals. Working with the posterior distribution is difficult in network models due to the combinatorial complexity and discrete nature of DAG-space. The performance of the proposed sampling scheme is also explore on synthetic data.

Systems genetics approaches to health and disease

Lars Steinmetz

EMBL Heidelberg

Universal therapies for disease are difficult to find, motivating the search for personal treatments that are tailored to the genetic constitution and environmental exposures of the patient. To achieve this goal, it is important to identify and intervene in the precise molecular pathways that are affected in a given patient and cause disease. I will present work in my laboratory that addresses these issues. We have been working on a recently discovered rare genetic disease caused by a recessive mutation in the NGLY1 gene. Besides common symptoms including developmental delay and movement disorders, affected children display a variety of more unique phenotypes such as the inability to produce tears. In the laboratory, we are applying transcriptomics and proteomics approaches directly to patient-derived cell lines in order to determine both common and individual effects of NGLY1 mutations, thus paving the way towards personalized therapies targeting this peculiar disease. In addition, I will present work on advancing transcriptome characterisation to the single cell level. We have developed technologies to measure heterogeneity in transcript isoforms and abundance between single cells and are applying these technologies to understanding immunity and host-pathogen interactions.

The use of hidden Markov models for the analysis of genomics and imaging data

Henrik Failmezger[1,2], *Sebastian Dümcke*[1,2], *Benedikt Zacher*[3], *Arijit Das*[1,2] *and Achim Tresch*[1,2]

[1] *Max-Planck Institute for Plant Breeding Research, Cologne, Germany*
[2] *Institute for Genetics, University of Cologne, Germany*
[3] *Gene Center, Ludwig-Maximilians-University Munich, Germany*

Introduction Current sequencing-based experimental techniques like RNA-seq and ChIP-Seq generate a wealth of data which can be aligned to the genome of the targeted organism. Yet their interpretation is not straightforward, and the size of the data alone requires automated, statistical analysis methods. Similarly, single cell time lapse imaging permits the investigation of dynamic perturbation responses at an unprecedented resolution. High throughput image analysis extracts morphological features pertaining to each cell image, resulting in a collection of genealogical trees where each node represents one cell at one time point by a high dimensional feature vector. In both situations, computational biology faces the challenge of grouping observations into functionally distinct classes while accounting for dependency between observations that is induced by the underlying linear or tree structure [1,2].

Results We present some hidden Markov models, which have been tailored to the above applications. We will highlight the problems that arise in their learning, derive algorithms for their solution, and show results obtained on experimental data. First, we introduce the bidirectional hidden Markov model (bdHMM, [3]), an HMMs satisfying a generalized version of the well-known detailed balance relation for reversible HMMs. BdHMMs are able to model DNA-related, forward and reverse strand-specific processes. They have been used to investigate the process of mRNA transcription, in which the Polymerase II transcription complex changes its composition several times along a transcript. We

will show how bdHMMs are employed to shed light on the "histone code" of the DNA, thereby creating an improved annotation of the human epigenome. Second, double-stranded HMMs (dsHMMs) are considered. They represent an alternative to the bdHMM by truly modeling two hidden Markov chains running in opposite direction. Third, we present the hidden Factor Graph model (HFM), a graphical model that can process tree-structured data (Fig. 1). We extend the theory of hidden Markov models to hidden variables whose dependence structure is encoded by a tree. We derive an efficient sum-product algorithm for the calculation of the marginal HFM likelihood. Based on this, we implement a Baum-Welch algorithm for parameter estimation. The HFM was applied to the Mitocheck database [4], a large compendium of time lapse movies of human cell lines in which single genes were knocked down by RNA interference.
We demonstrate that HFMs are able to

- Identify and characterize condition-specific abnormal morphological phenotypes
- Quantify condition-specific deviations from the normal dynamics of the cell cycle
- Quantify the contribution of stochasticity to cellular behavior

Literature

1. Zhong et al., Nat. Methods (2012) 9: 711-713.
2. Failmezger et al., BMC Bioinformatics (2013) 14.1: 292.
3. Zacher et al., Mol. Syst. Biol. (2014) 10: 768.
4. Neumann et al. Nature (2010) 464.7289: 721-727.

Talks

Genomics of ex-vivo drug sensitivity in primary tumour cells

Simon Anders, Leopold Sellner, Malgorzata Oles, Sascha Dietrich, Sophie Rabe, Wolfgang Huber and Thorsten Zenz

EMBL Heidelberg

New targeted therapeutics are currently opening new treatment options for many cancer types. Efficacy, however, is very variable across patients with the same cancer entity, and predicting to which drugs a patient's tumour will be sensitive is a major challenge.
We have conducted a drug screen on primary tumour cells from a cohort of leukaemia patients (chiefly CLL), testing ex-vivo the sensitivity of these patients' malignant cells to a panel of various drugs. Our data comprises, besides the screen output, further molecular characterizations of the samples by omics techniques (RNA-Seq, DNA methylation arrays, whole-exome sequencing) and by clinical assays.
Analysis of these data requires an integrative approach to find correspondences between the various data matrices, and so gain understanding of the molecular basis for the observed heterogeneity in drug responses. I will report on the methodological challenges we encountered and our approaches to address them.

Pascal: pathway scoring algorithm

Sven Bergmann

University of Lausanne

Integrating SNP p-values from genome-wide association studies (GWAS) across genes and pathways is a strategy to improve statistical power and gain biological insight. Here, we present Pascal (Pathway scoring algorithm), a powerful tool for computing gene and pathway scores from SNP-phenotype association summary statistics. For gene score computation, we implemented analytic and numerical solutions to efficiently derive test statistics. We examined in particular the sum and the maximum of chi-square statistics, which measure the strongest and the average association signals per gene, respectively. For pathway scoring, our method uses a modified Fisher method, which offers not only significant power improvement over more traditional enrichment strategies, but also eliminates the problem of arbitrary threshold selection inherent in any binary membership based pathway enrichment approach. We demonstrate the marked increase in power by analyzing summary statistics from a large number of different meta-analyses. Our extensive testing indicates that our method not only excels in rigorous type I error control, but also results in more biologically meaningful discoveries.

Analysis of thermal proteome profiling experiments

Dorothee Childs, Mikhail Savitski, Holger Franken and Wolfgang Huber

EMBL Heidelberg

Detecting the binding partners of a drug is one of the biggest challenges in drug research. Nevertheless, it is crucial to understand the drug's effects on the molecular level in order to infer its mode of action, as well as potential reasons for side effects. The recently introduced approach of Thermal Proteome Profiling (TPP) (Savitski et al, 2014) addresses this question by combining the cellular thermal shift assay concept with mass spec-trometry based proteome-wide protein quantitation. Thereby drug-target interactions can be inferred from changes in the thermal stability of a protein upon drug binding, or upon down-stream cellular regulatory events, in an unbiased manner. However, the analysis of TPP ex-periments requires a chain of novel data analytic and statistical modeling steps. Crucial steps involve the normalization over different replicates and treatment groups and the thermal sta-bility assessment of each protein. For this purpose, non-linear melting curves are fitted for each protein, yielding characteristic melting points for each treatment condition. Candidates with changed thermal stability under drug treatment are identified by detecting significant changes in the estimated melting points. To facilitate this process, we developed the TPP package for analyzing thermal profiling ex-periments. Here, we highlight the statistical challenges concerning the data processing and analysis and demonstrate how they are currently addressed in the package. Furthermore, we demonstrate how TPP experiments can help providing new insights in understanding drug effects. We hope that the availability of standardized and executable workflows will promote the adaptation of the powerful TPP method by the community, and aid Drug Discovery.

Literature Savitski, Mikhail M., Friedrich BM Reinhard, Holger Franken, Thilo Werner, Maria Fälth Savitski, Dirk Eberhard, Daniel Martinez Molina et al. "Tracking cancer drugs in living cells by thermal profiling of the proteome." Science 346, no. 6205 (2014): 1255784.

Cancer progression with Bayesian inference for acyclic digraphs

Jack Kuipers, Jonas Behr, Giusi Moffa and Niko Beerenwinkel

ETH Zurich

The progression of cancer may be viewed as an evolutionary process in which the accumulation of genomic alterations leads to certain capabilities of the cancer tissue that are essential for example to evade the host immune system and continue growth. Which genomic alterations are beneficial for tumour progression at a given point in time depends, however, on previous alterations and the environment of the cancer cell. Understanding these dependencies will improve cancer driver identification, elucidate mechanisms of resistance, and open new opportunities for personalised cancer treatment.

A number of models have been proposed to estimate these dependencies from cross-sectional data. Most of these (e.g. [1,2]) limit the structure of dependencies to trees and estimate the probability of an event given its single parent event. Conjunctive Bayesian Networks (CBNs) [3] consider a subset of directed acyclic graphs (DAGs) to model dependencies. The underlying assumption of CBNs is, that a mutation can only occur once all the mutations it depends on (parent events) have occurred.

We relax this strict assumption and for each gene we fit parameters for each state of its parent events over the complete space of DAGs. Other than for tree structures, inference of Bayesian networks is in general computationally demanding. For DAGs we utilise our novel algorithm based on their combinatorial structure [4]. This allows inference for larger gene networks than for previous methods for CBNs limited to about 15 nodes.

Therefore, more plausible biological models can be used to estimate likely progression paths for genetic data from large cohorts like the TCGA [5]. In addition to mutation data the model allows us to consider the influence of other factors like clinical data (treatment or stage, stroma fraction). We estimate progression models for the 12 main TCGA cancer types and analyse the similarities and differences of cancer progression across samples. Furthermore, we explore clustering of cancer patients based on similarities derived from our probabilistic graphical models.

Literature

1. N. Beerenwinkel, J. Rahnenführer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. Bioinformatics, 21:2106–2107, 2005.

2. R. Desper, F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. Journal of Computational Biology, 6:37–51, 1999.

3. M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel. Quantifying cancer progression with conjunctive Bayesian networks. Bioinformatics, 25:2809–2815, 2009.

4. J. Kuipers and G. Moffa. Partition MCMC for inference on acyclic digraphs. arXiv:1504.05006.

5. J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Mills Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. Nature Genetics, 45:1113–1120, 2013.

Removing batch effects and unwanted variation in gene expression data using nearest-neighbour matched sampling

Ivo Kwee[1,2,3], *Luciano Cascione*[1,3], *Andrea Rinaldi*[1], *and Francesco Bertoni*[1]

[1] *Institute of Oncology Research, Bellinzona, Switzerland*
[2] *IDSIA Dalle Molle Institute for Artificial Intelligence, Manno, Switzerland*
[3] *SIB Swiss Institute of Bioinformatics, Laussanne, Switzerland*

Statistical analysis of gene expression and other high-dimensional genomic data greatly suffer from batch effects and unwanted variation. These high-dimensional data are especially prone to small global changes that can obscure the actual variation of interest. The causes underlying unwanted variation are often not clear and they are a mixture of technical biases, biological variation and random noise. The increasing availability of public data sets enables the possibility for meta analysis and there is an interest to be able to correct data for unwanted variation to achieve better statistical power. First approaches to correct for batch effects in genomics data, such as quantile normalization, attempted merely to standardize the data in order to make samples more comparable. More recent approaches are including batch effects explicitly in the statistical model in order to more drastically reduce unwanted vari- ation. Lazar et al. divide the batch removal methods in two main approaches: location-scale (LS) methods and matrix-factorization (MF) methods [4]. The first class of method transforms the data to have similar mean and/or standard deviation, while the latter group of methods explicitly model the batch effects using factor analysis. Location-scale methods include batch mean-centering, gene-standardization, quantile nor- malization, ratio and scaling-based reference normalization methods that require reference samples, cross- platform normalization, empirical Bayes method as implemented in ComBat [3], and distance-weighted discrimination. Matrix-factorization methods include SVD/PCA-based batch effect substraction meth- ods such as surrogate variable analysis (SVA) [5] and the family of "remove unwanted variation" (RUV) methods [1, 2]. Here we present a new batch correction method for gene expression data based on nearest-neighbour matched sampling [6, 7]. The main idea of matched sampling is to pair each sample with similar samples but that have opposite label. By doing so, batch effects and individual bias are in many cases automatically corrected for and batch-induced variance is very much reduced. We compare our method with state-of- the-art methods, such as SVA, ComBat, and RUV.

Literature

1. J. A. Gagnon-Bartsch and T. P. Speed. Using control genes to correct for unwanted variation in microarray data. Biostatistics, 13(3):539–552, Nov. 2012.

2. L. Jacob, J. Gagnon-Bartsch, and T. P. Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. ArXiv e-prints, Nov. 2012.

3. W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1):118-27, Jan. 2007.

4. C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Sols, R. Duque, H. Bersini, and A. Now. Batch effect removal methods for microarray gene expression data integration: a survey. Briefings in Bioinformatics, 2012.

5. J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics, 3(9):1724-35, Sept. 2007.

6. D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association, 74(366):pp. 318-328, 1979.

7. E. A. Stuart and D. B. Rubin. Best Practices in Quantitative Methods, chapter Best practices in quasi- experimental designs: Matching methods for causal inference, pages 155-176. Sage Publications, 2008.

Cell type-specific regulatory circuits reveal modules disrupted in complex diseases

Daniel Marbach, David Lamparter, Zoltán Kutalik and Sven Bergmann

University of Lausanne

Mapping the molecular circuits that are perturbed by genetic variants underlying complex traits and diseases remains a great challenge. Here we integrate human promoter and enhancer activity data with transcription factor (TF) sequence motifs to infer a unique panel of $\sim$400 cell type and tissue-specific regulatory circuits. We find that similarity between these regulatory networks closely reflects developmental and functional relationships of corresponding cell types, tissues and organs. We further compile 37 genome-wide association studies (GWASs) and show that trait-associated genetic variants – including variants that do not reach genome-wide significance – perturb genes within cell type-specific regulatory modules. In most cases, evidence for perturbed modules is stronger than for protein-protein interaction networks and highly specific to disease-relevant cell types and tissues. Our results highlight the value of cell type-specific regulatory networks for understanding the genetic basis of complex diseases.

Estimation of a regulatory network of cooperation response genes in a model of cancer malignancy

Matthew N. McCall[1,2], Helene R. McMurray[2], Anthony Almudevar[1], and Hartmut Land[2,3]

[1] *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY*
[2] *Department of Biomedical Genetics, University of Rochester, Rochester, NY*
[3] *James P. Wilmot Cancer Center, University of Rochester Medical Center, Rochester, NY*

Advances in genomic technology have led to the discovery of numerous genes whose expression differs between cellular conditions; however, genes do not act in isolation, rather they act together in complex networks that drive cellular function. By considering the interactions between genes (and gene products), one gains a more in-depth understanding of the underlying cellular mechanisms. Estimation of these gene regulatory networks is necessary to understand cellular mechanisms, detect differences between cell types, and predict cellular response to interventions. Cancer progression has been shown to produce drastic changes in genetic networks critical to normal cellular function. Some oncogenic mutations produce self-sustaining alterations in the network structure such that removal of the original mutation does not restore normal cellular function. This suggests that identifying the original oncogenic mutation may not be sufficient for a targeted intervention; rather, a detailed understanding of the gene regulatory networks present in both normal and malignant cells may be necessary. Gene perturbation experiments are the primary tool to investigate gene regulatory networks and predict cellular response to interventions. However, given the numerous sources of variability and technical biases inherent in genomic technologies, network estimation algorithms are often unable to accurately reconstruct gene networks. To address this challenge, we propose an approach to network estimation that explicitly models and incorporates uncertainty in each step of the analysis. Instead of attempting to infer a single "best" network, we report a posterior density on the network space that directly conveys the uncertainty in the inferred network structure. Quantifying the uncertainty in specific network features allows researchers to determine highly-probable features and areas in which additional information is needed, thereby guiding future experimentation. We have applied the proposed model to a network of cooperation response genes (CRGs), which respond synergistically to loss-of-function p53 and Ras activation. CRGs have been shown to play a crucial role in tumor formation independent of the initiating mutations, and many CRGs are essential components of the cellular machinery involved in maintaining malignancy. Finally, there is evidence of a robust regulatory structure that governs patterns of CRG co-expression. Linking phenotypic variables (e.g. tumor growth), experimental perturbation of CRGs (via retrovirus-mediated re- expression of corresponding cDNAs or shRNA-dependent stable knock-down), and features of the CRG gene regulatory network provides a glimpse into the complex multi-gene regulatory relationships that are crucial to the malignant state. Ongoing examination of the CRG network architecture has the potential to uncover specific vulnerabilities of the cancer cell and, ultimately, to guide multi-target interventions.

The epigenetic code: experimental and computational approaches to unravel the interplay between multiple epigenetic and regulatory layers

Mattia Pelizzola

Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139 Milan, Italy

Epigenetic layers, namely DNA methylation and histone post-translational modifications, are recognized as a major player in cellular differentiation and increasing evidences confirm their involvement in various diseases, including cancer, thus making them suitable as biomarkers and drug targets. Despite the availability of vast large-scale public data, and the promotion of international consortia to profile epigenetic patterns in dozens of tissues and cell types, the interplay between epigenetic layers, their interaction with the regulatory machinery and the alteration of this crosstalk in cancer and other diseases remain fundamentally uncharacterized. In the next years we hope to contribute to this critical area of investigation by combining experimental and computational approaches in the following research areas:

1. Inference of tissue and disease specific (epi)genomics regulatory modules, based on the integrative analysis of large-scale public (epi)genomic datasets.

2. Studying the interplay between the RNA-stability machinery (based on 4sU-sequencing and RNA methylation experiments) and the epigenome

3. Deciphering the interplay between epigenetic and regulatory layers by precision epigenome engineering (such as CRISPR tools designed to specifically target the epigenome)

4. Studying epigenetic mechanisms controlling the interplay between transcription factors and the epigenome (adopting methods such as TRIP, to probe the epigenome by high-throughput integration of constructs including synthetic promoters upstream reporter genes)

Perturbation experiments are the key for most of these goals, as well as the development of computational methods for the analysis of the resulting data. Based on recent proof-of-concept studies from other groups and few preliminary results from my lab (for aims 1 and 2), I will discuss the intended research directions.

Modeling protein signaling pathways with Boolean Nested Effect Models

Martin Pirkl and Rainer Spang

University of Regensburg

Understanding cell signaling pathways is key in battling cancer. Extensions and revisions of current pathway models can have immediate consequences on treatment strategies. Here we introduce Boolean Nested Effects Models (B-NEM), a novel method to infer Boolean pathway structures as hypergraphs from gene expression data observed in per- turbation experiments. The pathway genes are perturbed alone and in combination. From the patterns of downstream effects in gene expression we infer the underlying pathway of signal transduction. Unlike existing Nested Effect Models the hypergraphs of B-NEM account for different modes of biological signal propagation like complex formation (AND-gate) in which all upstream players are necessary to transduct the signal and alternative signaling (OR-gate) in which only one active upstream player is necessary. Since the space for Boolean hypergraphs is large, a priori restrictions of the search space and efficient search methods are needed. Search space restriction is based on the biolog- ical literature or expert knowledge, while for the search algorithm, we propose a greedy neighbourhood search, which uses special features of Boolean algebra to escape greedily reached local minima.

Estimating causal effect strength between chromatin modifiers and histone modifications

Mohammad J. Sadeh and Martin Vingron

Max Planck Institute for Molecular Genetics, Berlin

Histone modifications and Chromatin modifiers are known to play an important role in the regulation of transcription. They are components of a chromatin-signaling network involved in transcription and its regulation. While individual modifications have received much attention in genome-wide analyses, little is known about causal interactions between chromatin modifiers and histone modifications. Here, we apply computational methods to recover causal effects be- tween chromatin modifiers and histone modifications from genome-wide ChIP-Seq data. We infer undirected networks based on partial correlations between histone modifications and chromatin modifiers. We then estimate the causal effects using the intervention calculus. We use summary measures of the set of possible causal effects to determine variable importance. Many recovered effects have literature support; others provide hypotheses about yet unknown causal relations. We show that the levels of chromatin modifying proteins predict histone modifi- cation levels and vice versa on a genome-wide scale in human cells. Our results suggest that our computationally derived effects are likely to lead to novel biological insights required to establish the connectivity of the chromatin-signaling network involved in transcription and its regulation.

Deconvoluting off-target confounded RNA interference screens

Fabian Schmich, Ewa Szczurek and Niko Beerenwinkel

ETH Zurich

Small-interfering RNAs (siRNAs) exhibit strong off-target effects, which confound the gene-level interpretation of RNAi screens and thus limit their utility for functional genomics studies. Here, we present gespeR, a statistical model for reconstructing individual, gene-specific phenotypes (GSPs). We demonstrate, that deconvolution of image-based infection screen phenotypes using over 115,000 siRNAs, single and pooled, from three companies and three pathogens, substantially improves the reproducibility between independent siRNA sets targeting the same genes. Genes selected and prioritized by our model were validated *in silico* and *in situ* and shown to constitute biologically relevant components of the entry mechanism of three different pathogens.

Inferring the underlying null model in three-way contingency tables with application to transcription factor co-occurrence

Alena van Bömmel and Martin Vingron

Max Planck Institute for Molecular Genetics, Berlin

In molecular biology, the ranked list representation of molecules or of experimental results is very common. The association of two ranked lists partitioned into two categories can be depicted by 2-way contingency tables. Introducing a third dimension (or a third ranked list) leads to a 3-way contingency table. While there is only one null hypothesis for the Fisher's exact test for the 2-way contingency table, there are several potential hypotheses to test in multiway contingency tables. The choice of a proper null model is enormously important since the underlying model influences dramatically the derived p-values and therefore the obtained results. However, the null hypothesis underlying the biological systems conditions is not always obvious.

Here, we present a rationale to determine the proper null model in 3-way contingency tables using the distribution of obtained p-values. For random data, one expects a uniform distribution of p-values. Thus for a mixture of random data and of a biological signal, one expects a slight enrichment of small p-values reflecting the non-randomness in the biological signal. Hence, we search for the model that is closest to such distribution.

We apply our approach to a problem of detecting transcription factor (TF) co-occurrence on tissue-specific promoters. The TFs are represented by a ranked gene list based on the binding affinity of the TF to the promoter sequences. The co-occurring TFs are those TF pairs that share a significant overlap of top-ranked target genes, additionally stratified by tissue. To determine the significance of this overlap leads to a 3-way contingency table test. Using the available data of predicted binding affinity for a large number of TFs on human promoters allows us the selection of the proper null model, which in this case is a model of partial independence. We believe that our strategy can be applied to other problems in biological systems such as perturbation studies. There, the relationship between perturbed variables stratified by cause or effect can be of interest. An analogous representation with ranked lists would also lead to a 3-way contingency table.

A multi-step classifier identifies cohort heterogeneity in cancers leading to improved accuracy of prognostic biomarkers

Jean Yee Hwa Yang

University of Sydney

The extensive genetic diversity in tumors may have a substantial effect on key cancer pathways and poses a significant challenge to personalized cancer medicine. Previous work has shown that individual patient samples may be considered as either "hard" or "easy" to classify by various clinical or molecular biomarkers, reflecting cohort heterogeneity that results in moderate error rates in prognostic outcome. Thus suggesting more complex models for prognostic classification are needed.

Here we describe the development of a two-step classifier that achieves dramatic and clinically relevant reduction in error rates. In the first stage of the classifier, which acts as a proxy for the identification of among-sample heterogeneity, we use clinico-pathologic variables to partition samples which can be explained by the gene expression data. Given this partition, we show that a second stage of analysis, modelling of the remaining samples with clinic-pathological variable, gives more accurate prognostic outcome compared with established one-step multivariate modelling of these features alone. We illustrate this "two-step" paradigm using three independent cohorts; melanoma, breast, and colorectal cancers where we observe 19%, 54% and 21% reductions in error rates respectively in comparison to the next best classifiers.

In statistical terms, our strategy models cohort heterogeneity via the identification of interaction effects in a high dimensional setting. At the translational level, this points to specific clinical attributes responsible for "hard" and "easy" classify samples and paves the way for a new generation of prognostic biomarkers for complex disease.

Capturing rewiring events during network evolution underlying dynamic biological processes

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

Cancer Research UK Cambridge Institute, University of Cambridge

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.

Multi-omics integration reveals local and distal genetic control of chromatin states and gene expression

Judith Zaugg

EMBL Heidelberg

Genome-wide association studies (GWAS) have uncovered thousands of correlations between genetic variants and complex human diseases. The majority of disease-associated variants, however lie in the non-coding part of the genome, which makes them difficult to interpret mechanistically. Therfore, to understand how natually occuring genetic perturbations, such as single nucleotide polymorphisms (SNPs), can affect gene regulation is fundamental to understanding human disease. Although gene regulation often involves long-range enhancer-promoter interactions it is unknown to what extent genetic variants in these elements act distally. Here we integrate chromatin profiling data for three promoter/enhancer histone marks in 76 individuals with HiC- and ChIA-PET-based physical chromatin interaction maps. We uncover interconnected networks of genetic links among regulatory elements: 10-15% have local histone quantitative trait loci (hQTLs), 15% of which affect distal elements. Physically interacting elements jointly contribute to gene expression, suggesting coordination among enhancers as well as genes. Transcription factor motif disruptions are enormously enriched in hQTL peaks for local and distal sites and alter gene expression over long distances. Importantly, hQTLs are enriched for immune disease and cancer GWAS-SNPs. Overall, we show that genetic variation affects networks of regulatory elements and sequence variation in these elements may play an important role in mediating phenotypic variation in humans.

Multi-omics data integration identifies causal mediators of global phenotypes

Chenchen Zhu

EMBL Heidelberg

A major challenge in systems genetics concerns the identification of causal intermediates underlying the genotype to phenotype path. This is a challenge because we lack a clear understanding of which genetic variants affect complex traits and how those effects are exerted at the molecular level. To unravel the chain of underlying biological events, intermediate traits such as gene expression levels have been used to establish links between genomic variation and global phenotype. As products of cellular pathways, proteins and metabolites represent promising candidates for unraveling intermediate cellular processes and reflecting the physiological state of a cell. In our pilot study, we have collected genome-wide gene expression profiles of a yeast cross between a laboratory strain and a clinical isolate of Saccharomyces cerevisiae in different environmental conditions. A statistical method has been developed to infer causal intermediate transcripts by exploiting environmental perturbations, thus distinguishing them from correlative effects downstream of phenotype. Currently, we are systematically collecting dynamic profiles of proteins and metabolites for the same panel of segregants. By taking advantage of the defined genetic perturbations presented in this population, the genetic regulation of each molecular layer is dissected using quantitative trait locus (QTL) mapping. We extend the existing causal inference method to the multidimensional dataset and exploit the relationships between the different molecular layers to learn biological principles that guide the conditioning of complex phenotypes. We present a new method to identify molecular signatures that are predictive for genes with a causal role in phenotype, integrating the proteomics and metabolomics data with existing genotypes, expression profiles, and growth rates for these strains. The integration of the proteomics and metabolomics data will permit the study of how genetic variations and transcript abundance impact cellular states. With this large compendium of datasets, we increase the sensitivity to detect causal networks at different levels. Beyond proposing a new route towards identifying specific molecular targets from high- throughput "omics" data, our results contribute to defining better models for how genotypic variation leads to phenotypic variation among closely related individuals.

Posters

Attractor landscape analysis through Boolean logical modeling of lung morphogenesis

Balaguru Ravikumar, Sarang Talwelkar, Jenni Lahtela, Jing Tang, Emmy Verschuren and Tero Aittokallio

University of Helsinki, Institute for Molecular Medicine Finland (FIMM)

Mouse lung morphogenesis is a highly dynamic process resulting from the crosstalk of signals between the epithelial and mesenchymal cells. This dynamic process extends across various phases of morphogenesis including initiation, elongation and termination phases. We aimed at improved understanding of this crosstalk system through modeling the underlying protein signaling network. Although there are various ways to model such a process, we have made use of deterministic Boolean logical models with synchronous updates. We specifically concentrated on the elongation phase of mouse embryo (E 9.5 – E 11), bringing insights into the critically of genes that are necessary for lung bud elongation and bifurcation

Large scale RNAi screens: analysis and correction of off-target effects

Neha Daga

University of Zurich

Small interfering RNAs (siRNAs) find wide application in high-throughput genome-wide functional screens. It has been described as well as observed that siRNAs which are designed to silence specific transcripts also down regulate various non-specific transcripts. This makes the interpretation of the screening data very difficult giving rise to many false positives.Therefore, a method was developed to address this issue. The method is based on seed phenotypes which are inferred from genome-wide data to correct the seed mediated off-target effects. We also review all the current methods available to correct the seed effects . These methods are applied to various genome wide pathogen screens and across different libraries .Various benchmarks are used to validate and compare these methods. One of them being counting the number of shared top hits between the libraries. This study is extensively performed to give an overview of how much one could correct the off-target effects and use the existing screens with least false positives before proceeding towards experimental validation.

Inferring modulators of genetic interactions with combinatorial Nested Effects Models

Madeline Dieckmann

ETH Zurich

Perturbation experiments have demonstrated many cases in which an organism can easily survive loss of a gene. This leads to the assumption of functional redundancies within a network. Genetic interaction maps can illustrate these redundancies, where interactions are defined as the difference between the combined phenotypes of two single-gene perturbations in contrast to the phenotype obtained from a double perturbation. However, these maps become difficult to interpret when one discovers differing epistatic interaction effects between two genes across varying gene sets. We show that a third gene can explain mixed epistasis, acting as a modulator of these interactions. In order to provide for genetic interactions, we extend the framework of Nested Effects Models by introducing logic gates to double perturbations. These logic gates represent interactions between regulators on downstream genes. We infer with high accuracy the right network from simulation studies. Furthermore, we apply our model to kinome perturbation data in S. cerevisiae and show that combinatorial NEMs can reveal modulators of genetic interactions.

Marginal integration for fully robust causal inference

Jan Ernest and Peter Bühlmann

Seminar for Statistics, ETH Zurich, Zurich, Switzerland

We consider the problem of inferring the total causal effect of a single variable intervention on a response variable of interest. We propose a certain marginal integration regression technique for a very general class of potentially nonlinear structural equation models (SEMs) with known structure, or at least known superset of adjustment variables and prove that it achieves the convergence rate as for nonparametric regression: for example, single variable intervention effects can be estimated with convergence rate $n^{-2/5}$ assuming smoothness with twice differentiable functions. This result can also be seen as a major robustness property with respect to model misspecification. We empirically compare the marginal integration regression method with more classical approaches and argue that the former is indeed more robust, more reliable and substantially simpler. Finally, we demonstrate a possible application of our methodology to gene expression data from the isoprenoid biosynthesis in Arabidopsis thaliana (Wille et al., 2004).

Distance correlation and its application in gene networks reconstruction

Mahsa Ghanbari

Max Planck Institute for Molecular Genetics

The assessment of relation between random variables is a common problem. For example, in gene networks reconstruction we are interested in finding the association or dependency among genes form gene expression data. While (partial) correlation is the most widely used method to measure the dependence between random variables, it captures only linear association and its power is reduced when associations are nonlinear. In addition, zero (partial) correlation means independency just in the case of Gaussian distribution. As a result, in many problems such as reconstruction of gene networks, where the nonlinear association between variables is common and the underlying distribution of data is far from being multivariate Gaussian, nonlinear dependence measurements that do not assume any distribution for the data can be more useful. Recently proposed method called distance correlation characterizes nonlinear dependence between random vectors. Distance correlation between two random variables is zero if and only if the random variables are statistically independent (without assuming any distribution for the data). The ability of distance correlation to identify nonlinear relationships between two random variables for many situations has been shown. However, in the multivariate analysis, such as gene networks reconstruction, it is important to account for the influence of other variables on the relation between two variables in order to distinguish between

direct and indirect relationship. Therefore, the concept of partial distance correlation has been suggested recently. We explored the behavior of (partial) distance correlation as an independence test in different aspects and used it in different methods for reconstruction of gene networks and compare it to other tests of independence. We used both simulated and real data to assess the performance of the methods.

Statistical modelling of PAR-CLIP data

Monica Golumbeanu

ETH Zurich

RNA-binding proteins are involved in a wide range of key molecular processes including gene expression regulation and RNA metabolism. Photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) is an experimental technique which aims to reveal the loci where a given protein binds to the RNA molecule. Built on next generation sequencing, the protocol adds supplementary binding evidence through induced T-to-C substitutions at the RNA-protein cross-link loci. The experiment is subject to noise and contamination and thus spurious T-to-C alterations are as well introduced. Therefore, PAR-CLIP data analysis requires appropriate pre-processing and thorough statistical modelling. We introduce BMix, a novel method based on probabilistic modelling of substitutions in PAR-CLIP data, which identifies and discards erroneous substitutions and reports high-confidence cross-link loci. Our method outperforms existing approaches both in accuracy and speed on real and simulated PAR-CLIP data.

A mathematical model of the evolution of tumor genomes during systemic tumor progression

Franziska Görtler

University of Regensburg

Systemic cancer progression is an evolutionary process. Selective pressures on cancer cells change during progression of the disease and the genomes of cancer cells have to adapt.[3,4] Different selective pressures in different microenvironment leads to an evolution in meta- stases different to the primary tumor. The instability of cancer genomes leads to a lot of mutations. Only the driver mutations affect cellular functions and give the cells a selective advantage for growth, disemination or survival in metatstatic niches. In breast cancer copy number aberrations (CNAs) are the predominant type of genomic alteration.[2] To model cancer progression I will use Conjunctive Bayes Networks (CBNs).[1] The goal in my thesis is to identify genomic changes in cancer cells which contribute to dissemination, survival and outgrowth in a metastatic niche. I have to develop a data pre-processing and modeling pipeline for array CGH data for modeling genomic cancer evolution during systemic progression of the disease. Up to now there exists no cancer progression model which includes metastases. The model shine light to the temporal order in which genomic aberrations occur during cancer development, dese- mination, parallel progression and outgrowth in different niches. The model will be applied to classify the driver alterations according to their role in systemic cancer progression. I have to fnd the driving CNA events and characterize them by their location/target (geno- me or signaling pathway) and type (deletion or amplification). Therefore I use CBNs. With the informations about the CNA events I will model systemic cancer progression. I will use CBNs to find sub-set structures in the binary data and the modeling of the systemic progression of the disease. The problem of long runtimes of CBNs with more than twelve nodes I want to overcome by the utilization of Approximate Bayes Computation methods. The cancer progression model shines light on the dissemination process and enables to classify the driver mutations by what they are driving.

Literature

1. Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. Conjunctive Bayesian networks. Bernoulli, 13(4):893-909, 11 2007.

2. Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. Genome Research, 22(2):398-406, 2012.

3. Vogelstein Bert and Kinzler Kenneth W. Cancer genes and the pathways they control. Nat Med, 10(8):789-799, aug 2004. 10.1038/nm1087.

4. R. Weinberg. The Biology of Cancer, Second Edition. Taylor & Francis Group, 2013.

Inference of signal transduction pathways from phosphorylation data to identify targets of combinatorial cancer therapy

Torsten Gross and Nils Bluethgen

HU Berlin

Over-activation of the MAPK/ERK signaling pathway can lead to uncontrolled cell growth and has been associated with many types of cancer. Detailed knowledge about the underlying network has therefore led to the discovery of potent treatments in targeted cancer therapy. However, intrinsic and acquired resistance through network rewiring corrupts their effectiveness, calling for combinatorial therapeutic interventions. The ability to design such appropriate treatments requires the identification and quantification of related kinase interactions. To this end, we conducted an experimental survey of phosphorylation states of selected kinases in different cancer cell lines and measured the response to various types of stimulations and inhibitions. However, these measurements represent the aggregate interplay between all involved kinases and do not directly quantify their direct interactions. To overcome this challenge, we propose a computational method that infers local interaction strengths between pairs of network components from global steady states. Its crucial advantage is a high practicality, as it allows for an efficient and coherent mathematical treatment of large networks, unobserved network components, and noisy data.

Empirical modelling of sequence space and fitness landscapes

Rasmus Henningsson

Lund University

The sequence space of viral quasispecies is of extremely high dimension, making it challenging to study and understand. We develop a sequence space model based on codon usage patterns across a viral genome. Dimension reduction techniques are then applied to reduce the dimension and assess how well the structure of a data set is captured in the low-dimensional representation. For several data sets, we show that the low-dimensional representation is able to find important features of the sequence space, allowing for an accessible visualization of the data. Furthermore, we describe how empirical fitness landscapes can be constructed, relating the low-dimensional sequence space position to measured fitness values. Finally, we evaluate the fitness landscape model using a data set that explores how robustness, the ability to buffer mutations, affects viral fitness.

Probabilistic approaches to estimate the mutation history of a tumor

Katharina Jahn

ETH Zurich

The recent development of single-cell sequencing techniques revealed that the genetic make-up of tumors is better described as a heterogeneous cell population than a monoclonal cell mass. Sequencing data of a large number of single cells can be used to reconstruct the evolutionary history of a tumor population. A key task in this procedure is to populate the ancestral states in the phylogenetic tree with mutations that split the cell population into subclones. Recurrent mutation orders observed in multiple tumor instances may lead to a better understanding of the mutational patterns associated with a specific tumor type, and help with the identification of tumor subtypes. Currently one of the major challenges in analyzing single cell sequencing data is its low quality due to the limited amount of DNA obtainable from a single cell. The main sources of error are a high allelic dropout rate and an increased false discovery rate compared to bulk sequencing. We introduce a probabilistic approach to estimate the mutation history of a heterogeneous tumor based on single-cell sequencing data that can deal with various error-types in the data. The method is evaluated in a simulation study and used to reconstruct mutation histories of different tumor types.

tba

Steffen Jaensch

Johnson & Johnson

In drug development, knowledge about the molecular mechanism of action (MoA) of small molecules is important for decision making but often not readily available. We aim to use a multi-parametric high-content screening approach to phenotypically classify compounds by screening test compounds along with target-annotated reference compounds. After extracting some hundred single-cell features by automated image analysis, which each give rise to a dose-response curve per compound, a phenotypic distance between each pair of compounds is computed. On the phenotypic distance we applied various clustering and dimensionality reduction algorithms to phenotypically group similar compounds. Notably, we found t-distributed stochastic neighborhood embedding (t-SNE) for rending a 2D map of the compounds that roughly reflects their high-dimensional distances, to be helpful for interpreting the results.
We have employed this approach to identify novel compounds that target the androgen receptor (AR) and inhibit its undesired activities in prostate cancer, including nuclear translocation leading to up-regulation of AR regulated genes. In addition to features related to prostate specific markers, we also extract information on nuclear and cellular morphology that can be utilized for identifying off-target toxicity effects. For several previously uncharacterized compounds we could predict and experimentally verify their mechanisms of action, including histone deacetylase (HDAC) inhibition and microtubule stabilization.

Improved correlation of genome-wide RNAi screens by accounting for seed-sequence effects

Alok Jaiswal, Jing Tang, Peddinti Gopalacharyulu, Krister Wennerberg and Tero Aittokallio

Institute for Molecular Medicine Finland (FIMM)

Genome-wide RNAi-based loss-of-function screens are routinely being used to probe functional dependencies of cells. However, these powerful tools have largely been shrouded in controversy due to lack of consistency between experiments. The sub-optimal reproducibility is largely due to off-target and other unpredictable factors. We compared two publicly available genome-wide shRNA screens performed on a compendium of cancer cell lines, and find that "seed-region" mediated sequence effects are consistent across identical cell lines screened in the two studies. Across 13 cell lines commonly screened in the Achilles study (Proc Natl Acad Sci U S A. 2011; PMID: 21746896) and Essential genes study (Cancer Discov. 2012; PMID: 22585861), we observed that the correlation between the activity of the shRNAs increases significantly when these data are analyzed by categorizing the shRNAs based on the sequence identity of the seed region (2-8 nt) of the shRNAs, rather than using the standard analysis based upon the identity of their genomic targets. Our results indicate that while the seed-region mediated off-target effects are ubiquitous amongst genome-wide RNAi screens, these may also be consistent across different studies. Taking into account such sequence effects can improve the reproducibility of the genome-wide shRNA screens in the future.

Comparison of methods for network reconstruction and their ability to identify predictive features

Jonatan Kallus[1], *Patrik Rydén*[2] *and Rebecka Jörnsten*[1]

[1] *Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg*
[2] *Department of Mathematics and Mathematical Statistics, Umea University*

We present a comparative investigation of biological network reconstruction methods' ability to infer different network attributes. When network models are applied for downstream analysis, e.g. survival models or subtype identification, different network attributes may be important in different applications. Thus, the optimal network model depends on the purpose of the model. Four popular methods (glasso, WGCNA, GeneNet and ARACNE) are here evaluated on simulated data with regard to four network attributes. Shell nodes are nodes that connect two or several network communities. We show that shell nodes are difficult to infer and argue that they are of biological interest. Ongoing research consists of validating results on real data, further study of the role of shell nodes and optimizing reconstruction methods' to infer them correctly.

Modularity and coevolution of penicillin-binding-proteins

George Kritikos

EMBL Heidelberg

Bacteria have been evolving for millions of years to occupy diverse niches but also to live in communities. Despite the wide range of niches and the many different ways bacteria adapt, a number of core processes are widely conserved among them. At the same time these core processes have to communicate with other processes, some of them being species-specific. Core processes frequently feature modular proteins to achieve these interconnections, which have arisen at different points in evolution.

In my work I am developing computational methods that will enable us to link gene interconnections within bacterial core processes to their evolutionary context, as well as discover new gene interconnections using genomic context information. Key to our approach is leveraging modularity in key proteins of core processes, such as Penicillin-Binding-Proteins (PBPs) in the cell wall biosynthesis process. After discovering modules (conserved regions or domains) in PBPs, we compared their phylogenetic distribution to that of all other proteins across species. Using this approach, we were able to predict known physical binding partners of PBPs.

Estimation of causal effects in high dimensional gene regulatory networks from a mixture of observational and intervention data

Gilles Monneret[1]*, Andrea Rau*[2]*, Florence Jaffézic*[2] *and Grégory Nuel*[1]

[1] *LPMA, UPMC, Paris*
[2] *GABI, INRA, Jouy-en-Josas*

The estimation of causal effects from gene expression data represents a challenging statistical problem, particularly as high-throughput experiments such as microarrays and RNA sequencing tend to include a limited number of biological replicates and a comparatively large number of genes of interest. Recent work (Rau et al., 2013) indicated that the estimation of causal effects in the framework of directed acylic Gaussian graphs can be greatly improved through the joint use of both observational (i.e., steady-state) and intervention (i.e., gene knock-outs and knock-downs) gene expression data. However, the currently proposed approach is limited to the inference of small networks. In this work, we propose an extension to this model to improve its performance in high dimensional settings through the addition of a ridge penalty and the possibility to incorporate the graph skeleton as a priori knowledge. We illustrate the benefit of these extensions in high dimensional settings, both on the quality of causal effect inference and computing time, using the DREAM 4 benchmark data as well as a set of simulations. Finally, we discuss the use of our model on a set of real gene expression data.

Inference of causal gene targets from expression data

Heeju Noh and Rudiyanto Gunawan

ETH Zurich

Elucidating the mode of action of drugs and the pathogenesis of diseases have received a growing interest in drug discovery field, specifically for applications in poly-pharmacology (Hopkins, 2008) and drug re-positioning (Ashburn and Thor, 2004). A common problem within this general topic is finding genes that are directly perturbed from expression data. For this purpose, a class of network analysis methods has been developed based on a kinetic modeling of the gene regulatory network (GRN), including network identification by multiple regression (NIR) (Gardner et al., 2003), mode of action by network identification (MNI) (Di Bernardo et al., 2005), and sparse simultaneous equation model (SSEM) (Cosgrove et al., 2008).

The most recent method SSEM was derived from a steady state linear dynamic system model of GRNs:

$$C = AC + P \tag{1}$$

where C is the matrix of log-2 fold change (FC) expression, A is the weighted connectivity matrix of the GRN, P denotes the input perturbation matrix, and n and m are the number of genes and samples, respectively. The direct gene targets are obtained in two steps. The first step involves the inference of the GRN connectivity matrix A based on the following linear regression:

$$C^T = C^T A^T + \epsilon \tag{2}$$

where ϵ denotes the noise term. Since the number of genes n is usually much larger than the number of samples m, a Lasso regularization is used by applying least angle regression (LARS) algorithm (Efron et al., 2004). In the second step, a ranking of candidate gene targets in a given drug treatment sample j is generated in the order of decreasing magnitudes of the residuals.

The predictions from SSEM therefore depend on the quality of the inferred GRN. In this work, we developed a method, called ΔNet, in which the gene targets are inferred directly from expression data. The method is based on solving Equation (1) as the following regression problem:

$$C^T = \begin{bmatrix} C^T & I \end{bmatrix} \begin{bmatrix} A & P \end{bmatrix}^T + \epsilon = XB + \epsilon \tag{3}$$

where I is the identity matrix. Thus, the gene targets P and connectivity matrix A are inferred simultaneously. We also employ LARS algorithm to estimate B. Despite the larger dimension of B in comparison to A, ΔNet analysis in the case studies finished sooner than SSEM.

We tested the performance of ΔNet in predicting the direct targets of test sets for E. coli and yeast. We gathered E. coli and yeast microarray datasets, consisting of 524 and 570 samples, respectively, from GEO, ArrayExpress, and M3D (http://m3d.mssm.edu) databases. We used samples in the dataset with known perturbations, e.g. knock-out and overexpression experiments (104 from E. coli and 157 from yeast), to evaluate and compare the performance of ΔNet, SSEM and Z-test (Tibshirani and Hastie, 2007). The results demonstrated that the direct gene targets are ranked higher by ΔNet than by SSEM and standard Z-test.

We also used ΔNet to analyze time-series expression data from human lung cells Calu-3 infected with influenza (90 samples, from the Kawaoka group at University of Wisconsin-Madison). For the analysis, we also compiled additional Calu-3 expression data from ArrayExpress database (564 samples). Finally, we performed Gene Ontology (GO) and pathway enrichment analyses (https://toppcluster.cchmc.org/) on top gene targets predicted by ΔNet (using $P > 0.5$). The GO analysis of ΔNet gene targets indicated that cellular processes associated with viral replication and defense response to virus, including interferon and cytokine signaling, are activated in early time points (within 7 hours post infection). The enrichment analysis further suggested human influenza as the most relevant disease. For later time points (beyond 12 hours), we observed much more differentially expressed genes, but the GO analysis of ΔNet gene targets suggested non-specific cellular response. Such insight is important in understanding the dynamics of cell response to influenza viral infection, which may lead to novel therapeutic intervention.

Literature

1. Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews. Drug Discovery, 3, 673-683.

2. Cosgrove, E. J., Zhou, Y., Gardner, T. S., & Kolaczyk, E. D. (2008). Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. Bioinformatics, 24(21), 2482-2490.

3. Di Bernardo, D. et al. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nature Biotechnology, 23(3), 377-383.

4. Efron, B. B., Astie, T., Johnstone, I., & Tibshirani, R. (2004). LEAST ANGLE REGRESSION. The Annals of Statistics, 32(2), 407-499.

5. Gardner, T., Bernardo, D. Di, Lorenz, D., & Collins, J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. Science, 301(July), 102-106.

6. Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. Nature Chemical Biology, 4(11), 682-690.

7. Tibshirani, R., & Hastie, T. (2007). Outlier sums for differential gene expression analysis. Biostatistics, 8(1), 2-8.

Causal structure search in bow-free acyclic path diagrams

Christopher Nowzohour

ETH Zurich

Structure learning is one of the main objectives of statistical causal inference. While good methods exist for the case when all variables are observed (and the corresponding model has a DAG representation), this is not the case for more general model classes. We present a greedy algorithm for structure learning with bow-free acyclic path diagrams (BAPs), which can be viewed as a generalization of directed acyclic graph (DAG) models. Additionally, we investigate distributional equivalence properties of these models and use these results to quickly discover causal effects from an identified BAP. The application of our method to various real-world datasets reveals that BAP models can represent the data much better than DAG models in most cases.

Investigating eQTL role in inter-individual variation of transcriptomic response to bacterial infection

Barbara Piasecka

Institut Pasteur, France

The human immune system keeps us in generally good health. However, personal susceptibility to infections and disease severity is highly variable among individuals and may lead to different immune dysfunctions. For example, one can develop inflammation, autoimmunity disorder, allergy or even cancer. Similarly, the response to the therapy and its side effects occurrence is different between humans. But medical research and healthcare system typically ignore individual heterogeneity and take "one size fits all" approach both for disease management and for drug development. The Milieu Intérieur Project aims to characterize a healthy response of an immune system, and to discover the genetic and environmental factors involved in naturally present variability. To this end, a total of 1000 healthy French volunteers split equally by sex and stratified across five-decades of life were recruited. Every person has been examined in the context of i) every-day life habits (through an extensive questionnaire); ii) genomic variability (with genome-wide SNP genotyping); iii) metagenomic diversity (with sequencing of microbial populations in fecal and nasal samples); iv) induced transcriptional and protein signatures (with whole blood stimulation system); and v) variability in levels of circulating immune cell populations (based on flow cytometry). Here, I focus on variation in gene expression across individuals, which has been suggested to be the determinant of phenotypic variation and susceptibility to disease. In particular, I will present some preliminary results in exploring the association between genomic variability and variation of transcription signatures induced by three different bacterial stimulations of subjects' immune systems (E. coli, S. aureus, BCG).

Integrating multiple Omics data in a gene-network oriented view

Wolfgang Raffelsberger

CNRS UMR 7104-DIVISION 1051 IGBMC

In overall we are interested in combined analysis different high throughput screening methods to gain insight into pathological or beneficial effects of novel medical treatments. In particular, we are studying means of enhancing oncolytic activity of Vaccinia viruses towards solid tumors. To this purpose a genome-wide RNAi screen accompanied by RNAseq was performed and is getting complemented by proteomics data. While initial data analysis uses current platform-specific approaches and algorithms, the final integration of gene-product linked toxicity or viability with gene-product abundance estimates is primarily based on a permutation testing strategy and is compared with simple intersection of platform independent tables. Finally, all results are reconstituted and visualized in a protein interaction (PPI) network context (Intact, String) to identify and rank patters with common central regulator-genes or events. In overall, we were able to identify gene-products either promoting cancer cell survival or enhancing virus toxicity. However, the elevated number of "isolated" entry points raises the hypothesis that PPI based maps should be more dynamic and may be at the present state oversimplifying the complexity of the regulatory processes involved in committing to cell death or survival.

Complex systems in animal health

Simon Rüegg[1], *Laura Vinckenbosch*[2], *Katharyn Mitchell*[3], *John Berezowski*[4] *and Christian Mazza*[2]

[1] *Section of Epidemiology, Vetsuisse-Faculty, University of Zurich*
[2] *Department of Mathematics, University of Fribourg*
[3] *Clinic for Equine Internal Medicine, Vetsuisse-Faculty, University of Zurich*
[4] *Veterinary Public Health Institute, Vetsuisse-Faculty, University of Bern*

In human and veterinary medicine, complex system theory can, in principle be applied to systems of any size ranging from systems of molecules, to cells and even to large populations of individuals. The theory suggests interesting questions when confronted with ill health at any of these scales. We are applying the theory to a number of cases in order to explore potential uses for complexity metrics in medical applications. We have identified Composite Multiscale Sample Entropy (CMSE), Correlation Dimension (CD) and Resilience derived from the fluctuation-dissipation-theorem as metrics which may find application in clinical settings. All three can be computed from time series data that is becoming increasingly available in everyday practice. We will present the results of computing these measures in two data sets. 1) A group of 40 cows, (30 lame and 10 not lame), which were fit with pedal accelerometers for several days. We were blinded to the lameness status of the cows and challenged to distinguish between the two cow groups using these metrics. 2) Electrocardiographic (ECG) recordings from 3 healthy horses and 2 horses with arrhythmias. ECG recordings were taken while the horses were at rest, during periods of exercise on a treadmill at different intensities and immediately after exercise. The ECG time series were evaluated using these metrics to determine if there are measureable differences between periods of rest versus exercise, versus recovery and also to determine if and how resilience, i.e. the return to baseline heart rate following a perturbation (exercise) may be affected by arrhythmias. Furthermore, we will assess the effect of exercise and arrhythmia on the sample entropy and the correlation dimension.

Machine learning aided neuroradiological diagnostics of brain tumors

Johannes Slotboom[1], *Martin Zbinden*[1], *Nicole Porz*[2], *Alessia Pica*[3], *Roland Wiest*[1], *Stefan Bauer*[1,4], *Andeas Raabe*[1], *Philippe Schucht*[1] *and Beate Sick*[5,6]

[1] *DRNN-DIN/SCAN University Hospital Berne, Berne, Switzerland*
[2] *Institute of Neurosurgery, University Hospital Berne, Berne, Switzerland*
[3] *Institute of Radiooncology, University Hospital Berne, Berne, Switzerland*
[4] *Institute of Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland*
[5] *Division of Biostatistics, ISPM, University Zürich, 6 IDP ZHAW*

Magnetic Resonance Imaging (MRI) is often used for the initial diagnosis of brain tumors. However, despite the great expert knowledge of the involved medical doctors and neuroradiologists the differential diagnosis for individual patients remains often difficult. To support this process we train a random forest (RF) with features extracted from recorded MR images and corresponding confirmed diagnoses. As data basis we have used apparent diffusion coefficient maps (ADC-maps) which is the most commonly used type of MR-imaging and does not require injection of MR-contrast agent. We have chosen the RF as machine learning algorithm since it is known for its high accuracy especially in situations with many interacting and possibly noisy features. Here we demonstrate how detailed texture features beside the common intensity means and additional patient characteristics as input to a RF classification algorithm yield tumor class predictions with an accuracy that is valuable to the specialist physician during the differential diagnosis.

Gene Networks from time series with spike-and-slab feature selection

Edgar Steiger

Max Planck Institute for Molecular Genetics

Time series data of the dynamics of gene expression becomes increasingly available and allows the inference of causality and the reconstruction of gene regulatory networks. Reliable algorithms are needed to reveal these causal relationships between genes as well as the appropriate delays of effects from the noisy data. I implemented the algorithm CATSS built on a linear regression model with a time series matrix from the data. A Bayesian approach took care of the need for sparsity by imposing a Categorical distribution in conjunction with a Spike-and-Slab prior on the regression coefficients. The parameters were iteratively matched to the data with the Expectation Propagation algorithm. CATSS revealed itself to be a fast and dependable alternative to existing alternatives, showing greater accuracy in comparisons on synthetic and real data. This method exhausts the capabilities of a linear model to reconstruct networks and is able to derive and even disclose possibly unknown gene-specific delays.

Automatic identification of the regulators mediating the indirect signal from transcription factor to the target genes

Elena Sügis, Andreas Ellervee, Hedi Peterson and Jaak Vilo

Institute of Computer Science, University of Tartu, Liivi2, 50409,Tartu, Estonia

To the date large amount of studies have addressed the questions of identifying the regulators of certain biological conditions from large-scaled experiments. However, different types of experiments reveal only a part of the real regulatory relationships between genes and proteins in the system. The combination of the variety of experiments could more accurately describe the direct and indirect relationships that determine the status of the cells in a particular biological condition. Perturbation experiments allow to understand which genes are affected in a certain biological condition after knock-down or over-expression of a given gene. These effects may have direct or indirect nature. Combining perturbation results with other types of data, such as chromatin immunoprecipitation, could help to distinguish the genes that are directly or indirectly affected. In this work we combine transcription factors' perturbation experiments with their chromatin immunoprecipitation data sets for automatic identification of the regulators mediating the indirect signal from transcription factor to the target genes. We apply our pipeline on the human embryonic stem cells data from ESCDb (http://biit.cs.ut.ee/escd/) and ENCODE project data repository (https://www.encodeproject.org/search/?type=experiment).

Signaling pathway models from perturbation data

Ewa Szczurek

ETH Zurich

Signaling pathways constitute the machinery underlying basic cellular processes, and their malfunction gives rise to diseases. Pathogens, for example, can manipulate the signaling networks of their hosts in order to invade the target cell. Cancer is driven by multiple genomic alterations that target key oncogenic pathways and alter the way they work. Our goal is to reproducibly model signaling pathways and their influence on observed phenotypes from perturbation data, where genes are silenced or removed, and phenotypic effects of these interventions are measured. Several obstacles stand on the way to fulfill this goal.

First, in most experimental setups, perturbations target the layer of the signaling pathway, whereas the phenotypes are measured on a distinct layer downstream, connected by an unknown intermediate layer, which complicates learning predictive models of the system. Second, the standard experimental perturbation technique, RNA interference with silencing RNA (siRNA), is confounded by so-called off-target effects that are due to additional, sequence-dependent targeting of multiple genes, other than the intended on-target. As the third obstacle, siRNA knock-downs do not have 100% efficiency and dampen each of their on- and off- targets expression with a different strength. Both the unspecific targeting and its variable strength can cause the often reported un-reproducibility of siRNA screens between independent labs, and should be taken into account together with the way such perturbations may propagate in the pathway.

I will propose an approach to model pathways from perturbation data, based on three realistic assumptions. First, I assume that perturbations propagate in the pathway via the signaling interactions. In other words, I expect that with a perturbation targeting a node in the pathway with a certain strength, also other nodes reachable from it in the pathway are perturbed. Second, I assume that each perturbed gene has its individual contribution to the observed phenotype. Finally, I assume that the phenotypes are a weighted sum of the individual gene contributions, with weights set to their perturbation strengths. In this framework, siRNA on- and off-targeting is treated as combinatorial pathway perturbations for enhancement of model inference.

Design of experiments for causal inference of gene regulatory networks

S. M. Minhaz Ud-Dean and Rudiyanto Gunawan

ETH Zurich

In this study, we addressed the problem of inferring gene regulatory networks (GRNs) from expression data of gene perturbation experiments. Such inference is known to be very challenging, primarily because the problem is underdetermined. Typically there exist not one, but an ensemble of network structures that are consistent with the data. For this reason, we previously developed an ensemble inference algorithm, called TRaCE (Transitive Reduction and Closure Ensemble) [1], which generates the upper (most complicated network) and lower (least complicated network) bounds of such an ensemble. Applying TRaCE to in silico generated data from random and benchmark GRNs demonstrated the high importance of gene perturbation experiments from which the data came, in determining whether the inference has a unique solution. Despite the wide recognition of the underdetermined nature of GRN inference, design of experiment (DOE) for such an inference has received much less attention than the development of new inference methods. Our work here seeks to fill this gap.

In this work, we represented GRNs as directed graphs (digraphs), where the nodes represent genes and the (directed) edges represent gene regulatory interactions. We developed the DOE based on ensemble inference and graph theory. Specifically, we use the upper bound and lower bound of the ensemble from TRaCE to determine uncertain edges, defined as edges in that do not belong to . We introduced the concept of edge separatoid, which is a set of genes in the GRN that by knocking them out, could lead to the verification of an uncertain edge. In the DOE, we employed genetic algorithm to find the optimal set of genes whose knock-out (KO) could potentially verify the maximum number of uncertain edges. This DOE is then iterated with TRaCE, where and are continually updated with new experimental data of the optimal KO set from the previous iteration. We first tested the iterative GRN inference procedure, combining the DOE and TRaCE, for the inference of E. coli GRN. Here, we assumed that we could obtain the verifiable edges of a given set of KO genes without error. The iterative procedure converged to the true GRN of E. coli in 154 iterations when excluding KOs of essential genes and limiting the number of gene KO to up to 20 genes in the experiments. In contrast, and from the complete set of single- and double-gene KO (DKO) data, also assuming ideal condition as above, did not give a unique GRN solution. We also applied the DOE to in silico generated steady state gene expression data using a stochastic differential equation (SDE) model of GRN. The SDE model was simulated using the Euler-Maruyama method [2] and the data were then contaminated with 5% log-normal random noise. We performed the iterative procedure for five benchmark 100-gene networks from DREAM4 in silico network inference challenge [3]. The results summarized in Table 1 show that the iterations of DOE and TRaCE could give a single GRN solution with only a small number of errors.

Literature

1. Ud-Dean, S.M.M. and R. Gunawan, Ensemble Inference and Inferability of Gene Regulatory Networks. PLoS ONE, 2014. 9(8): p. e103812.

2. Higham, D.J., An algorithmic introduction to numerical simulation of stochastic differential equations. . SIAM review, 2001. 43(3): p. 525-546.

3. Schaffter, T., D. Marbach, and D. Floreano, GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics, 2011. 27: p. 2263-2270.

Outlier detection in Cox proportional hazards models based on the concordance c-index

João Diogo Pinto[1]*, Alexandra M. Carvalho*[1,2] *and Susana Vinga*[3]

[1] *PIA, Instituto de Telecomunicações, Lisboa, Portugal*
[2] *DEEC, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal* [3] *IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal*

Survival analysis allows to build regression models and classifiers using time-to-event or lifetime data, identifying relevant features associated with the outcome. Cox proportional hazards model is probably the class of models most applied, where the baseline hazard function is multiplied by a linear combination of the covariates. The robustness of the Cox regression has been proven to be rather weak to the presence of outlying observations, which can severely affect the coefficients estimation. Outlier detection, in this context, is a relevant task to identify observations that do not follow the overall pattern of the model and might be related with interesting characteristics that can be further explored. We propose two outlier detection methods for survival data under the Cox model using the Harrell's concordance c-index as a measure of goodness of fit. The c-index is related with the predictive power of a regression by assessing if higher risks correspond to lower survival times. The first method is a single-step procedure that presents a delete-1 statistic based on bootstrap, testing for the increase in the concordance c-index. The second method is based on a sequential procedure that maximizes the c-index of the model using a greedy one-step- ahead search. Finally, we use both methods to perform robust estimation for the Cox regression, removing from the sample the most outlying observations. Our preliminary results on one simulated and two real datasets illustrate the improvement of the estimation of the regression coefficients and also the increase of Cox model predictive ability.

Genome-scale Mapping of Signaling Networks Underlying Cell Migration

Franzisca Völlmy

Biotech Research and Innovation Center (BRIC) University of Copenhagen (UCPH)

Despite the recognized importance of cellular migration and its role in the pathological behavior of cells, the signaling dynamics as well as the underlying genetic events controlling the cellular phenotype lack comprehensive characterization at a systemswide level. We aim to uncover the signaling dynamics involved in coherent cell migration by

studying two breast cancer cell lines MCF7 and MDAMB231, respectively nonmigrating and migrating, and their response to wounding. Within the time span of the woundclosing process, quantitative global massspectrometry based phosphoproteomics data will be acquired. Morphological information is provided by a genomewide siRNA screen for realtime screening of processes associated with live cell migration. These proteomic, phenotypic, as well as sequencing data are integrated into a predictive model with the aim of gaining a deeper understanding of migration and its involvement in cancer. Finally, this approach should lay the groundwork for identification of potential future drug targets against invasive phenotype in cancer.

Asymptotic inference for integral curves of noisy vector fields

Susan Wei and Victor Panaretos

EPF Lausanne

Noisy vector fields arise in many disciplines of study. One of the most fundamental operations on a vector field is its integration. We derive conditions under which the integral curve of a noisy vector field is asymptotically normal. Results in this vein have previously been obtained in the specific contexts of diffusion tensor imaging in neuroimaging and filament estimation in astrophysics. The theory we build decouples the analysis from the specific vector field estimator used. Numerous illustrative examples will be given to demonstrate the wide- ranging applicability of our results.

Participants

Rita Achermann University of Zurich
Simon Anders European Molecular Biology Laboratory (EMBL)
Brenda Andrews University of Toronto
Niko Beerenwinkel ETH Zurich
Roderick Beijersbergen The Netherlands Cancer Institute
Diego Benusiglio University of Turin
Sven Bergmann University of Lausanne
Blake Borgeson Recursion Pharmaceuticals
Michael Boutros DKFZ Heidelberg
Peter Bühlmann ETH Zurich
Anne Carpenter Broad Institute of Harvard and MIT
Luciano Cascione Institute of Oncology Research
Anatole Chessel University of Cambridge
Dorothee Childs EMBL Heidelberg
Neha Daga University of Zurich
Madeline Diekmann ETH Zurich
Maria Pamela Dobay Swiss Institute of Bioinformatics
Anna Drewek ETH Zurich
Jan Ernest ETH Zurich
Bernd Fischer German Cancer Research Center (DKFZ)
Mahsa Ghanbari Max Planck Institute for Molecular Genetics
Monica Golumbeanu ETH Zurich
Franziska Görtler University of Regensburg
Torsten Gross HU Berlin
Rudiyanto Gunawan ETh Zurich
Rasmus Henningsson Lund University
Susan Holmes Stanford University
Wolfgang Huber EMBL
Nikolaos Ignatiadis EMBL - Huber Group

Steffen Jaensch Johnson & Johnson
Katharina Jahn ETH Zurich
Alok Jaiswal University of Helsinki
Vinay Jethava ETH Zurich
Jonatan Kallus University of Gothenburg
George Kritikos EMBL Heidelberg
Jack Kuipers ETH Zurich
Ivo Kwee Insitute of Oncology Research
Shu Li ETH Zurich
Marloes Maathuis ETH Zurich
Daniel Marbach University of Lausanne
Matthew McCall University of Rochester
George Michailidis University of Michigan
Gilles Monneret UPMC Paris
Heeju Noh ETH Zurich
Simon Flyvbjerg Nørrelykke ETH Zurich
Christopher Nowzohour ETH Zurich
Gregory Nuel CNRS
Mattia Pelizzola Istituto Italiano di Tecnologia (IIT)
Barbara Piasecka Institut Pasteur
Martin Pirkl University of Regensburg
Michael Prummer ETH Zurich
Wolfgang Raffelsberger CNRS UMR 7104-DIVISION 1051 IGBMC
Mattias Rantalainen Karolinska Institutet
Balaguru Ravikumar University of Helsinki
Hubert Rehrauer ETH / University of Zurich
Simon Rüegg University of Zurich
Hélène Ruffieux EPFL
Mohammad Sadeh Max Planck Instigate tute
Fabian Schmich ETH Zurich
David Seifert ETH Zurich
Beate Sick ZHAW
Edgar Steiger Max Planck Institute for Molecular Genetics
Lars Steinmetz EMBL Heidelberg
Elena Sügis University of Tartu
Ewa Szczurek ETH Zurich
Christian Tischer EMBL Heidelberg
Achim Tresch University of Cologne

S. M. Minhaz Ud-Dean ETH Zurich
Alena van Bömmel Max Planck Institute for Molecular Genetics
Susana Vinga IDMEC
Franziska Voellmy DTU
Susan Wei EPFL
Jean Yang University of Sydney
Vicente Yepez Ludwig-Maximilians Universität
Ke Yuan University of Cambridge
Judith Zaugg EMBL Heidelberg
Chenchen Zhu EMBL Heidelberg