

8th International Workshop on Bioinformatics and Systems Biology



 **Genomics and Systems Biology
of Molecular Networks**
International Research Training Group

*Berlin, Teikyo Hotel
June 9-11, 2008*

PROGRAM

Monday

8:45 – 9:00 am **Opening remarks**

9:00 - 10:00 am **Satoru Miyano**
Peta Flops Computing for Systems Biology

Signalling (chair: Knapp)

10:00 - 10:20 am **József Bruck**, Wolfram Liebermeister and Edda Klipp
The Effect of Variable Enzyme Concentrations in a Kinetic Model of Yeast Glycolysis

10:20 - 10:40 am **Alexander Skupin** and Martin Falcke
The Role of IP₃R Clustering in Calcium Signalling

10:40 - 11:10 am *Coffee Break*

11:10 - 11:30 am **Euna Jeong**, Masao Nagasaki and Satoru Miyano
Rule-Based Reasoning for System Dynamics in Cell Systems

11:30 - 11:50 am **Kaname Kojima**, S. Imoto, T. Shimamura, A. Fujita, S. Miyano
Estimation of non-linear Gene Regulatory Networks via L₁ Regularized NVAR from Time Series Gene Expression Data

Data technology (chair: Herzel)

11:50 am -12:10 pm **Max Flöttmann**, J. Schaber, E. Klipp, S. Hoops, P. Mendes
ModelMage: A Tool for the Automatic Generation and Discrimination of SBML-Models

12:10 - 12:30 pm **Raymond Wan**, Asa M. Wheelock and Hiroshi Mamitsuka
A Framework for Determining Outlying Microarray Experiments

12:30 - 2:00 pm *Lunch Break*

Saccharomyces cerevisiae (chair: Holzhütter)

2:00 - 2:20 pm **Clemens Kühn**
Exploring the Impact of Osmoadaptation on Glycolysis using Time-Varying Response-Coefficients

Networks

- 2:20 - 2:40 pm **David K. Byrne**, Timothy S. Gardner, Daniel Segrè
A Computational Framework for Microbial Nutrient Optimization
- 2:40 - 3:00 pm **Moritz Schütte** and Oliver Ebenhöf
A Network Evolution Model
- 3:00 - 3:20 pm **Kai Kruse**
Comparing Flux Balance Analysis to Network Expansion: Producibility, Sustainability and the Scope of Compounds
- 3:20 - 3:50 pm *Coffee Break*
- 3:50 - 4:10 pm **Timothy Hancock**, Hiroshi Mamitsuka
Multiclass Semi-Supervised Graph Partitioning with Decision Trees
- 4:10 - 4:30 pm **Jorge Numata**, Oliver Ebenhöf and Ernst-Walter Knapp
Measuring Statistical Coupling in Metabolic Networks Using Mutual Information
- 4:30 - 4:50 pm **Evan Snitkin** and Daniel Segrè
Optimality Criteria for the Prediction of Metabolic Fluxes in Yeast Mutants
- 4:50 - 5:10 pm **Thomas Handorf**
Nutritional Requirements of Metabolic Networks
- 5:10 - 5:30 pm **Yugo Shimizu**, M Hattori, S. Goto and M. Kanehisa
Generalized Reaction Patterns for Prediction of Unknown Enzymatic Reactions
- 5:30 - 7:30 pm *Poster session*

Tuesday

Bioinformatics (chair: Ebenhöf)

- 9:00 - 9:20 am **Mike Cui**, T. Smith, P. W. Robbins, and J. Samuelson
Darwinian Selection for Sequons (Sites of Asn-Liked Glycosylation) with Thr in Phylogenetically Disparate Eukaryotes and Viruses
- 9:20 - 9:40 am **Raphael A. Bauer**, K. Rother, J. M. Bujnicki and Robert Preissner
Suffix Techniques as a Rapid Method for RNA Substructure Search
- 9:40 - 10:00 am **Stephen C. J. Parker**, Elliott H. Margulies, and Thomas D. Tullius
The Relationship between Fine Scale DNA Structure, GC Content, and Functional Elements in 1% of the Human Genome
- 10:00 - 10:20 am **Jin Hwan Do**, Satoru Miyano
*Secondary Metabolite Gene Cluster Analysis by the Cumulative GC and DNA Curvature Profile in *Aspergillus fumigatus* Genome*
- 10:20 - 10:50 am *Coffee Break*
- 10:50 - 11:10 am **Yosuke Hatanaka**
A Novel Strategy to Search Concerted Transcription Factor Activities Using Gene Expression Profile and Genomic Data
- 11:10 - 11:30 am **Jonas Maaskola**
How important is sequencing accuracy for next generation "deep" sequencing?
- 11:30 - 11:50 am **Christian Waltermann**, M. Krantz, S. Hohmann, E. Klipp
*Sensitivity-Driven Model Discrimination in the HOG Pathway in *S. Cerevisiae**
- 12:30 pm *Lunch Break*

Trip to Spreewald (see information at the end of the book) and BBQ in Zeuthen

Wednesday

- 9:00 - 10:00 am **Morihiro Hayashida**
Domain-Based Model for Protein Interaction Network
- Gene regulation (chair: Falcke)*
- 10:00 - 10:20 am **Lan Hu**, Daniel Segrè, Temple F. Smith
Understanding the Regulatory Mechanisms of Retrotransposed Genes in the Drosophila genus
- 10:20 - 10:40 am **Manuela Benary**
Modeling the Gene Expression of IL-2 in Single T Cells
- 10:40 - 11:10 am *Coffee Break*
- 11:10 - 11:30 am **Marta Luksza**
Cross-species analysis of gene expression
- 11:30 - 11:50 am **Swantje Struck**, U. Schmidt, B. Grüning, R. Preissner
SuperToxic: a Systems Biology Approach to Toxicity
- 11:50 am - 12:10 pm **Marvin Schulz**, Edda Klipp
Identification of Optimal Drug Targets in Differential Equation Networks
- 12:10 - 12:30 pm **Masataka Takarabe**, S. Okuda, M. Itho, Toshiaki, S. Goto, and M. Kanehisa
Network analysis of adverse drug interactions
- 12:30 - 2:00 pm *Lunch Break*
- (chair: Günther)*
- 2:00 - 2:20 pm **Aysam Gürler** and Ernst-Walter Knapp
Sampling Protein-Protein Geometries in Real Space
- 2:20 - 2:40 pm **Stephan Lorenzen** and Ernst-Walter Knapp
Fully Flexible Refinement of Docking Decoys
- 2:40 - 3:00 pm **Benjamin Menküc** and Christoph Gille
Computer Aided Selection of Isotopomer Labels for Tracer Experiments
- 3:00 h - 3:20 pm **Christoph Gille**, Andreas Hoppe and Hermann-Georg Holzhütter
Publishing Annotated Sequence and 3D-Alignments
- 3:20 - 3:40 pm **Gyan Bhanot**, H. Liu, G. Alexe, D. Juan, T. Antes, C. Delisi, and L. Liou
MicroRNA Diagnostic Panels and Gene Targets in ccRCC
- 3:40 pm **Closing remarks**

Abstracts for the talk session

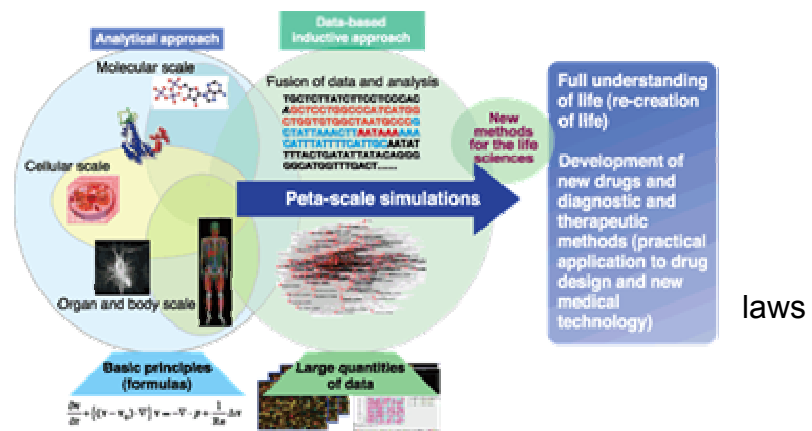
Peta Flops Computing for Systems Biology

Satoru Miyano

Human Genome Center, Institute of Medical Science, University of Tokyo

RIKEN has started the project for RIKEN Next-Generation Supercomputer R&D Center in 2006 and it is developing a supercomputer system with 10 peta flops computing ability in Kobe. Simultaneously, it launched the “Grand Challenge” project which will develop software applications for life sciences and nanotechnology. The grand challenge project for life sciences called “The Next-Integrated Life Simulations” aims at developing software applications that will enable us to simulate and analyze the processes that take place in living organisms, from the molecular level to the level of the whole body.

It takes two approaches. The first is an analytical approach, where biological/physiological phenomena (molecular scale, cellular scale, organ and body scale) are studied through basic principles (formulas and models). The second is a data-based inductive approach, where we will attempt to discover new processes and by analyzing large quantities of experimental data. We are involved with the second



approach. Obviously, biology and medicine are facing with large-scale high dimensional heterogeneous data, e.g. transcriptome, proteome, P-P interactions, SNPs, physiological data, diseases, phenotypes, etc. Moreover, currently investigated biological systems are incomplete and complex, and there are big gaps between models and real data. These models and data may be divided into two categories. One is the “general” one as *Homo sapiens*. The other is the “personal” one as an individual. Our aim is to bridge and fuse the gaps between “general” and “personal” by constructing a peta flops computational strategy that will develop and employ the following technologies:

- (1) Inferring and analyzing large-scale gene networks
- (2) Large-scale protein-protein interaction predictions
- (3) Large-scale SNP data analysis
- (4) Data assimilation for biological systems

With this strategy, we will contribute to practical application to new drug target discovery/design and diagnostic/therapeutic methods.

The Effect of Variable Enzyme Concentrations in a Kinetic Model of Yeast Glycolysis

József Bruck¹, Wolfram Liebermeister² and Edda Klipp²

¹ Chair of Theoretical Biophysics, Humboldt University Berlin

² Max Planck Institute for Molecular Genetics, Berlin, Germany

Metabolism is one of the best studied and understood fields of biochemistry, but its regulation involves processes on many different levels, some of which are still not understood well enough to allow for quantitative modelling and prediction. Glycolysis in yeast is a good example: although high-quality quantitative data are available, well-established mathematical models usually only cover direct regulation of the involved enzymes. The effect of various metabolites on the enzyme kinetics is summarised in carefully developed mathematical formulae. However, this approach implicitly assumes the enzyme concentrations themselves are constant, thus neglecting other regulatory levels – e.g. transcriptional and translational regulation – involved in the regulation of enzyme activities. It is believed, however, that different experimental conditions result in different enzyme activities regulated by the above mechanisms. Detailed modelling of all regulatory levels is still out of reach since some of the necessary data – e.g. quantitative large scale enzyme concentration data sets – are still lacking or rare. Nevertheless, a viable approach is to include the regulation of enzyme concentrations into an established model and investigate whether this improves the predictive capabilities. Proteome data are usually hard to obtain, but levels of mRNA transcripts may be used instead as clues for changes in enzyme concentrations.

Here we investigate the question whether and how far including mRNA data into an established model of yeast glycolysis allows to predict the steady state metabolic concentrations for different experimental conditions. To this end, we modified an established ODE model for the glycolytic pathway of yeast [1] to include changes of enzyme concentrations. The changes are inferred from mRNA transcript level measurement data [2]. We investigate how this approach can be used to predict metabolite concentrations for steady state yeast cultures at five different oxygen levels ranging from anaerobic to fully aerobic conditions.

References

[1] Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, Snoep JL: **Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry.** *Eur J Biochem.* 267(17) (2000), PMID: 10951190

[2] Wiebe MG et al: **Central carbon metabolism of *Saccharomyces cerevisiae* in anaerobic, oxygen-limited and fully aerobic steady-state conditions and following a shift to anaerobic conditions** *FEMS Yeast Research* 8(1) (2008)

The Role of IP₃R Clustering in Calcium Signalling

Alexander Skupin and Martin Falcke

Max-Delbrück-Center, Robert-Rössle-Str. 10, 13125 Berlin, Germany

Ca²⁺ is the most important second messenger controlling a versatility of intra and intercellular physiological processes. Most of these processes are controlled by oscillations of the cytosolic Ca²⁺ concentration as e.g. gene expression. These oscillations occur by the interplay of Ca²⁺ release from the endoplasmic reticulum (ER) into the cytosol via channels and the re-uptake of Ca²⁺ into the ER by SERCA pumps. A common channel type present in many cell types is the inositol trisphosphate receptor (IP₃R), which is activated by IP₃ and Ca²⁺ itself. This property leads to Ca²⁺ induced Ca²⁺ release (CICR), since Ca²⁺ released from one channel diffuses to adjacent channels and can open them, too.

The CICR was extensively studied in experiments with *Xenopus oocytes* demonstrating the spatial character of CICR based the fact that IP₃R channels are localized in channel clusters consisting up to 40 channels, which are separated within the cell between one to seven μm, i.e. in the order of one to two diffusion lengths. This inhomogeneous distribution causes a hierarchical development of Ca²⁺ spikes from local events, so called puffs, which might nucleate to a global wave, what is known as puff to wave transition. In other cell types with typically smaller cell sizes such transitions were not observed leading to the assumption of diffusive arranged channels, what would shrink the spatial character of Ca²⁺ signals. The main question is therefore, how essential the spatial effect of CICR is in general.

Experimental investigations are restricted to fixed arrangements of IP₃Rs, since IP₃R expression and diffusion occur on slower time scales as typical experiments and are for this reason not able to clarify the influence of IP₃R clustering. On the other hand theoretical studies have to deal with a stiff system of coupled partial differential equations driven by the stochastic channel behaviour. It turned out, that straight forward methods as Finite Element solver are too slow to solve the reaction diffusion system (RDS) describing the biological system.

We developed a method dealing sufficiently with both, the experimental set up of cells and the complex mathematical system by splitting the problem into two parts. For this purpose we derived an analytical solution of the corresponding RDS in terms of coupled Green's function and couple the local stochastic channel behaviour via a Gillespie algorithm to the global cell dynamics. Thus, we are able to study the spatially resolved cell behaviour for different arrangement of the IP₃Rs and can analyze the effect of IP₃R clustering to answer the question about the spatial character of CICR.

Rule-Based Reasoning for System Dynamics in Cell Systems

Euna Jeong, Masao Nagasaki, Satoru Miyano

Human Genome Center, Institute of Medical Science, University of Tokyo

We have investigated how logic-based rules can be used for OWL-based ontologies to infer new or implicit information and ensure consistency from existing knowledge. Using a system-dynamics-centered ontology developed for biological pathways in cell systems with its repositories, we have inferred the existence and type of biological entities participated in biological reactions. Additionally, we have regulated pathway models to reflect molecular activities actually occurred in living organisms and made them to be able to simulate based on a mathematical model called hybrid functional Petri net with extension. Through some examples, we show that the proposed rule-based approach helps researchers to explore numerous dynamic modelling and simulation tasks with no prior knowledge.

Estimation of Non-Linear Gene Regulatory Networks via L_1 Regularized NVAR from Time Series Gene Expression Data

Kaname Kojima, Seiya Imoto, Teppei Shimamura, André Fujita, Satoru Miyano

Human Genome Center, Institute of Medical Science, University of Tokyo

Recently, non-linear vector autoregressive model (NVAR) based on Granger causality was proposed to infer non-linear gene regulatory networks from time series gene expression data. Since NVAR requires a large number of parameters due to the basis expansion, and the length of time series data is insufficient, we need to limit the size of the gene set strongly. To address this limitation, we employ L_1 regularization technique to estimate NVAR. Under L_1 regularization, genes that are direct parents in the regulations can be selected efficiently even when the number of parameters exceeds the number of data samples. By comparing to the existing researches, we can more accurately estimate gene regulatory networks considering a larger gene set than those ones. Through the simulation study, we verify the effectiveness of the proposed method by comparing its limitation in the number of genes to that of the existing NVAR. The proposed method is also applied to time series microarray data of Human cell line.

ModelMage: A Tool for the Automatic Generation and Discrimination of SBML-Models

Max Flöttmann¹, Jörg Schaber¹, Edda Klipp¹, Stefan Hoops², Pedro Mendes^{2,3}

¹Max Planck Institute for Molecular Genetics, Berlin, Germany

²Virginia Bioinformatics Institute, VBI, VA, USA

³Manchester, UK

Objective

Mathematical modeling of biological systems involves implementing, testing and discriminating between model alternatives that differ in the number of components, reactions and/or kinetics. Generating and managing these model alternatives is a tedious and difficult task. ModelMage is a management tool that facilitates handling of candidate models. It is designed for the easy and rapid development, generation, simulation and discrimination of model alternatives. Here, we present the novel, yet simple, core technology and algorithm of ModelMage.

Results

The main idea of the program is to create a defined set of model alternatives in an automatic way. The user provides only one SBML-model and a set of directives from which alternatives are created by leaving out components and/or reactions. After generating the models, the software can automatically fit all these models to data and provide different statistical measures for goodness of fit to make discrimination between the models easier. In contrast to other model generation programs, ModelMage aims at generating only a limited set of models instead of all possible ones. Moreover, it uses COPASI as a simulation engine. Thus, all simulation and optimization features of COPASI are readily incorporated. We apply ModelMage to model alternatives of an artificial biochemical network and demonstrate how the correct model can easily be recovered.

Conclusions

ModelMage is a novel and efficient technology to generate families of models and rank them by goodness of fit to given data. During tests with artificial as well as real world models from the literature, ModelMage proves as a useful tool that could speed up modelling work substantially.

A Framework for Determining Outlying Microarray Experiments

Raymond Wan, Åsa M. Wheelock, Hiroshi Mamitsuka

Microarrays are high-throughput technologies that allow researchers to assess the expression levels of thousands of genes simultaneously. It is well-known that the data obtained from microarrays are noisy and may contain many missing values. Various efforts have been spent on addressing this noise through techniques such as normalization. However, if enough samples are available, the best technique in dealing with noisy microarray data is to simply employ more replicates.

In a set of k replicate experiments, it is possible that some are more problematic than others. While increasing the number of replicates would minimize the negative effects from these problematic replicates, we describe techniques for determining which *experiments* (instead of individual expression levels) are potential outliers.

The aim of this work is to establish techniques that would allow researchers to determine which experiments are questionable in an impartial way. Ideally, such techniques would be applied after the image acquisition phase but before any analysis, including averaging across experiments, is done.

The methods that we have considered range from minimizing a correlated profile based error function for expression level *cleaning* to the application of distance based outlier techniques from the field of data mining. The underlying theme is our attempt to use the entire experiment to assess its validity, rather than just focusing on a small subset of expressed genes (as one would for data mining purposes). We report on methods that have not worked, as well as ones which we believe still hold promise by applying a series of preliminary experiments.

Exploring the Impact of Osmoadaptation on Glycolysis Using Time-Varying Response-Coefficients

Clemens Kühn, E. Petelenz, B. Nordlander, J. Schaber, S. Hohmann, E. Klipp

The aim of this work is to investigate the mechanisms of glycerol accumulation in *Saccharomyces cerevisiae*, particularly the impact of increased glycerol production during osmotic adaptation on glycolysis. In response to hyperosmotic stress, baker's yeast accumulates glycerol as a compatible solute. Glycerol production is not exclusively regulated by osmotic adaptation, but also influenced by glycolytic activity and redox balance. We present a mathematical model based on existing experimental data as well as on theoretical work, which focuses on the interactions between osmoadaptation and glycolysis, namely the production of glycerol and its influence on flux towards pyruvate. Our model shows that increased glycerol production can have a substantial negative effect on the pyruvate production rate. The influence strongly depends on the initial balance between these processes. The model indicates a crucial role of the interaction between active Hog1 and Pfk26, which is a stimulator of glycolysis. This indication is in accordance with the existing experimental data (Dihazi et al. 2004).

A Computational Framework for Microbial Nutrient Optimization

David K. Byrne, Timothy S. Gardner, Daniel Segrè

Bioinformatics Program, Boston University, Boston, MA

Metabolic engineering in microbial hosts for the production of renewable chemicals and energy sources has received considerable attention in recent years. Microbial fuel cells (MFCs) and biodiesel, as representative forms of renewable bioelectricity and biofuel sources, provide new opportunities for the sustainable production of energy from biodegradable, reduced compounds. Yet MFC electrical current and biodiesel fatty-acid synthesis still fall short of the output and efficiencies desirable for practical implementation. Since microbial metabolism is the primary cellular mechanism by which energy is generated and distributed, it is of interest to develop methods that will increase the proportion of energy diverted for bioenergy use. In this study, a computational framework has been developed to design a nutrient optimization program that maximizes output and efficiency using models of microbial metabolism. Two optimization strategies were implemented. First, all possible nutrient combinations were analyzed. Second, using the strong duality theorem from linear programming, a bilevel optimization procedure was developed to generate nutrient compositions that optimally regulate the cell. The procedure ensures that any increase in the energy consumed to produce biomass is consequently coupled to bioenergy output yields. Optimization predictions indicate that significant improvements in *Shewanella oneidensis* MFC electricity generation and *Escherichia coli* biodiesel fatty-acid synthesis are attainable.

A Network Evolution Model

Moritz Schütte and Oliver Ebenhöf

The representation and analysis of huge amounts of biological data measured by high-throughput techniques poses major challenges in the field of theoretical biology. One way of illustrating the data is to represent it as a correlation network, such as metabolic networks where two metabolites (nodes) get connected if their concentration measurements from different experiments are highly correlated.

Similarly, in gene co-expression networks, nodes correspond to genes which are connected by an edge whenever the genes are similarly expressed over a variety of conditions and developmental stages. Such a co-expression network has been derived for the model plant *Arabidopsis thaliana* based on microarray data from 356 different experiments collected from the PubMed Gene Expression Omnibus (GEO) database. The resulting network contains 22810 nodes representing the *Arabidopsis* genes. Interestingly, this network displays a degree distribution deviating from a scale-free distribution in that it exhibits a sharp cut-off for high degrees and two humps just before the cut-off.

We have developed a model for network evolution which is not directly based on graph theoretical properties. Genes, represented as nodes, are described as D-dimensional unit vectors, characterizing their expression profile obtained from D experiments. Genes are linked if their Pearson correlation exceeds a certain threshold. Our model is initialized by a small number of genes, representing a small set of ancestral genes present during early stages of evolution. New genes appear by two mechanisms: single gene duplication or whole genome duplication. Gene loss is incorporated indirectly by assuming that old genes possess a higher chance of successful duplication and persistence in the genome. Mutations are introduced by small alterations of the D-dimensional gene expression profile vectors. We investigate how different parameters such as mutation rate or the number of whole genome duplication events influence graph theoretical properties of the resulting co-expression network. We can show that for a wide range of parameters our model represents characteristic features of the experimentally determined co-expression network for *Arabidopsis thaliana*.

By a systematic analysis how varying parameters influence characteristic properties of the resulting co-expression network, we expect that our model contributes to our ability to interpret organism-specific co-expression networks in the light of their evolutionary history.

Comparing Flux Balance Analysis to Network Expansion: Producibility, Sustainability and the Scope of Compounds

Kai Kruse

The producibility of metabolites from available resources is investigated systematically using flux balance analysis (FBA) and network expansion. Calculations are performed for the genome-scale metabolic networks of *Escherichia coli* and *Methanosarcina barkeri*. Strict biological interpretation of the results obtained with FBA leads to the concept of 'sustainability', which reduces the set of producible metabolites by assuming a growing and dividing cell. By applying this concept it can be shown that the resulting sets of producible metabolites from FBA and network expansion coincide in a large majority of cases. The purely heuristic approach of allowing for certain cofactors to facilitate reactions during the process of network expansion helps to refine those results even further. In conclusion, we state that network expansion, due to its enormous advantages in computational speed, is a valuable alternative to determining producible metabolite with FBA.

Multiclass Semi-Supervised Graph Partitioning with Decision Trees

Timothy Hancock¹, Hiroshi Mamitsuka¹

¹Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan

In this paper we investigate semi-supervised graph partitioning using decision trees to search for sub-graphs within a graph adjacency matrix. Graph partitioning by a decision tree seeks to optimize a specified graph partitioning index e.g. ratio cut by recursively applying decision rules found within nodes of the graph. Key advantages of tree models for graph partitioning is that they provide a predictive framework for evaluating the quality of the solution, determining the number of sub-graphs and assessing overall variable importance. We evaluate the performance of tree based graph partitioning on a benchmark dataset for multiclass classification of cancer diagnosis based on gene expression. Three graph cut indices will be compared, ratio cut, normalized cut and network modularity in terms of their predictive accuracy, optimal sub-graph number extraction and their relative power to extract know important feature within the dataset.

The benchmark dataset used for the comparison of the three metrics is the Ramaswamy et al. microarray dataset [1] for multiclass classification. This is an ideal benchmark as it is an combination of multiple mircoarray datasets each on a different tissue type and with a separate tumor and normal classification problem. This allows for two different scales for the structure of the same dataset. Firstly the large scale resolution is to classify tumor/normal irrespective of tissue type and the secondly the small scale structure is to classify tissue type irrespective of tumor/normal classification. The results in these two settings highlight the differences between the three graph cut indices and furthermore show that decision tree graph cutting is both an accurate and informative method of clustering.

References

[1] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. **Multiclass cancer diagnosis using tumor gene expression signatures**. *Proc Natl Acad Sci U S A*, 98(26):15149–15154, December 2001.

Measuring Non-Linear Correlation in Metabolomic Networks Using Mutual Information

Jorge Numata, Oliver Ebenhöf, Ernst-Walter Knapp

Macromolecular Modelling Group, Dept. of Chemistry and Biochemistry, Freie Universität Berlin, Takustr. 6,
Berlin 14195 Germany

A non-linear correlation coefficient based on mutual information is used to gauge the statistical dependency among sets of variables. While the Pearson correlation coefficient is only exact for linear dependency, the mutual information coefficient is applicable to other cases. The superiority of mutual information has been known for some time, but stable numerical implementations had been missing. Here, we use recent non-parametric algorithms based on k-nearest neighbour distances. The statistical significance of the results is probed for the relevant case of tens to hundreds of metabolomic data points using constructed examples. Finally, an application to experimental metabolite concentrations is shown.

Optimality Criteria for the Prediction of Metabolic Fluxes in Yeast Mutants.

Evan Snitkin and Daniel Segrè

Bioinformatics Program, Boston University 44 Cummington St., 02215, Boston

Constraint-based models of cellular metabolism, such as flux balance analysis (FBA), use convex analysis and optimization to study metabolic networks at a genome scale. The availability of reaction lists for numerous organisms, along with a variety of network analysis and optimization tools, is making these approaches increasingly popular for metabolic engineering and biomedical applications, as well as for addressing fundamental biological questions. It is therefore very important to assess the predictive capacity of these models and to understand how to interpret them in a biologically relevant manner. Typically, model assessment is limited to gauging the ability to predict phenotypes, such as viability under different environmental and genetic conditions. These types of assessments, for the most part, focus only on the growth phenotype of the cells, but ignore the underlying flux predictions. While this may be sufficient for certain types of study, the question of whether flux balance models can reliably predict intracellular and transport fluxes is crucial for more detailed analysis, and remains largely unanswered.

Here we compare FBA model predictions of yeast metabolic fluxes to a previously published set of experimentally determined fluxes for 14 different single gene deletion mutants across a variety of possible objective functions. We find that the specific optimization criteria used to determine fluxes have a significant impact on the accuracy of the predicted fluxes. For virtually all mutants, the best flux predictions were not generated using the commonly used objective of growth maximization, but by using a variant of the approach of minimization of metabolic adjustment (MOMA), in which we calculated the minimal flux rerouting from an experimentally constrained wild type flux prediction. Interestingly, while different optimization methods provide very different levels of agreement relative to experimental fluxes, they tend to provide similar predictions for viability. This demonstrates that assessment of models at the level of flux predictions is a critical step in assessing the biological validity of different models and optimization criteria.

Nutritional Requirements of Metabolic Networks.

Thomas Handorf

The main purpose of metabolism is to provide the necessary chemical precursors for almost all processes in the cell. Nutrients taken up by an organism are converted into the desired target metabolites through a network of metabolic reactions. The actual composition of the required set of nutrients depends on the capabilities of those reactions which in turn are defined by the enzymes encoded in the organism's genome.

In this work, the nutritional requirements of the metabolic networks of a vast number of organisms are systematically investigated. A set of target metabolites is defined, containing among others amino acids, energy equivalents like ATP, sugars, organic acids and metabolic cofactors. For each metabolic network, various possible nutrient sets are tested whether they can synthesize the desired target metabolites. These tests are performed using the concept of scopes [1]. As a result, for each organism several minimal nutrient sets are obtained.

The nutritional requirements of all organism specific networks were categorized in resource types. These types include sugars, organic acids, amino acids as well as vitamins like niacin, thiamine. Each organism can be assigned a nutritional profile indicating which of the resource types are essential, optional or not used in its metabolism.

The results [2] clearly distinguish between generalists like *E.coli* which only require a small set of essential nutrients and non-generalists, including parasitic species like *Buchnera*. A detailed analysis for *Buchnera* showed a good agreement of the predicted profile with experimental results.

The method presented can be used to predict possible growth media for cultivation of certain organisms or suggest possible vitamin requirements. This may also be of interest for the analysis of mutants, indicating the appearance of new metabolic deficiencies. Further, the prediction of nutrients for which no experimental verification is found can uncover possible shortcomings of the utilized metabolic networks. This information can in turn be used to enhance the annotation of these networks.

References

- [1] **Expanding Metabolic Networks: Scopes of Compounds, Robustness and Evolution**, Handorf, T., Ebenhöf, O., Heinrich, R., *J. Mol. Evol.*, 61:498-512, (2005)
- [2] **An environmental perspective on metabolism**, Handorf T., Ebenhöf O., Christian N., Kahn D., *Journal of Theoretical Biology*, in print (2007)

Generalized Reaction Patterns for Prediction of Unknown Enzymatic Reactions

Yugo Shimizu, Masahiro Hattori, Susumu Goto and Minoru Kanehisa

Bioinformatics Center, Institute for Chemical Research, Kyoto University

Prediction of unknown enzymatic reactions is useful for understanding biological processes such as degradation of proteins and reactions to external substances like endocrine disrupter. To improve the accuracy of the prediction, we need to define more reasonable similarity measure in the reaction. We have developed the KEGG RPAIR database which is a collection of chemical structure transformation patterns, called RDM patterns for substrate-product pairs of enzymatic reactions. In this study, we compared RDM patterns with EC numbers which are the well-know hierarchical classification scheme for enzymes, based on overall reactions. Additionally, we performed hierarchical clustering of RDM patterns using the information whether each sub-subclass of EC has the RDM pattern or not. To represent the variation of RDM patterns in a cluster, we generalized RDM patterns in the same cluster using the hierarchy of KEGG atomtypes which are the components of RDM patterns. Using this generalized pattern, we can predict which cluster includes a given RDM pattern even if the reaction of the pattern has not been assigned any EC numbers. Thus we will be able to define the similarity between enzymatic reactions by using this cluster information.

Darwinian Selection for Sequons (Sites of Asn-Liked Glycosylation) with Thr in Phylogenetically Disparate Eukaryotes and Viruses

Jike Cui, Temple Smith, Phillips W. Robbins, and John Samuelson

Numerous protists and rare fungi have truncated Asn-linked glycan precursors and lack N-glycan-dependent quality control (QC) systems for glycoprotein folding in the ER. Here we show that the abundance of sequons (Asn-Xaa-Thr or Asn-Xaa-Ser), which are sites for N-glycosylation of secreted and membrane proteins, varies by more than a factor of four among phylogenetically diverse eukaryotes based upon a few variables. There is positive correlation between the density of sequons and the AT-richness of coding regions, although no causality can be inferred. In contrast, there appears to be Darwinian selection for sequons containing Thr, but not Ser, in organisms that have N-glycan-dependent QC of glycoprotein folding. Selection for sequons with Thr, which nearly doubles the sequon density in human secreted and membrane proteins, occurs by conditional selection, wherein the actual sequon density is greater than the sequon density calculated from the frequencies of Asn, Thr, and Pro. Conditional selection also appears to account for increasing sequon densities of the haemagglutinin of influenza viruses A/H3N2 and A/H1N1 over the past few decades of human infection. Very strong selection for sequons with both Thr and Ser in gp120 of HIV and related retroviruses results from conditional selection for sequons, amino acid composition bias, and increases in AT-richness. In summary, AT-content is an important predictor of sequon density, conditional selection for sequons with Thr occurs in phylogenetically diverse eukaryotes with N-glycan-dependent QC of glycoprotein folding, and multiple mechanisms may contribute to the very high density of sequons in viral envelope proteins.

Suffix Techniques as a Rapid Method for RNA Substructure Search

Raphael A. Bauer^{1,2}, Kristian Rother³, Janusz M. Bujnicki^{3,4} and Robert Preissner¹

¹ Institute of Molecular Biology and Bioinformatics, Structural Bioinformatics Group, Charité Universitätsmedizin (Medical University), Arnimallee 22, 14195 Berlin, Germany

² Graduate School: Genomics and Systems Biology of Molecular Networks, Monbijoustr. 2, 10117 Berlin, Germany

³ International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland

⁴ Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

The RNA Ontology Consortium recently proposed a two-letter representation of RNA backbone conformations [1]. In total, 46 distinct conformations were identified from the torsion angles of C4'-C4' "suite" units.

In this study, we compare the suite notation to a custom string representation that utilizes eta-theta pseudotorsion angles. Both pseudotorsion angles were split into 36 ten-degree bins which can be converted into an alphabetical representation.

The suite representation as well as the converted pseudotorsion angles are therefore translations from 3D space into a 1D string. This enables us to use classical string matching methodology to detect structural features in turn.

To detect similarities between two RNA structures we are utilizing suffix techniques that allow to detect substructure similarity within some degree of inexactness. The analysis we present here was carried out using suffix arrays and suffix trees that have different scaling behavior towards runtime and memory consumption. A main feature of suffix trees and suffix arrays is that a substring search can be performed in a very fast manner against a huge set of strings. Specifically in $O(m)$ for suffix trees or $O(m \log n)$ for suffix arrays, with m being the length of the search string and n the amount of entries in the database.

Both representations were used to assess similarity and self-similarity in several RNA structure datasets.

The suite- as well as the pseudo-torsion representation was tested on

- 1) All motifs in the SCOR database.
- 2) The manually refined RNADB2005 dataset.
- 3) All high-resolution RNA 3D structures available from the PDB.
- 4) A set of 74 tRNA structures from all kingdoms of life.

The possibility to detect structural similarities on these datasets allowed to recover many homologous structural elements that have implications for further understanding of the RNA apparatus in Systems Biology.

We identified several examples of local structural variations in the suite representation, while the tertiary structure of the backbone is conserved. The mismatch-tolerant suffix string implementation and the pseudo-torsion representation are both able to recognize these. In many cases, even homologous motifs result in different suite representation. Non-homologous SCOR motifs are in most cases completely different.

We therefore conclude that the suite representation is too detailed to represent the backbone trace in a coarse-grained manner. On the contrary, the pseudo-torsion representation of the RNA backbone is able to recognize a wider range of variations by adjusting the mapping between character and the representing angle.

The software as well as the utilized data sets are freely available from <http://suiterna.sourceforge.net>.

References

[1] Jane Richardson et al. 2008. **RNA backbone: Consensus all-angle conformers and modular string nomenclature.** *RNA*;14(3):465-81.

The Relationship between Fine Scale DNA Structure, GC Content, and Functional Elements in 1% of the Human Genome

Stephen C. J. Parker¹, Elliott H. Margulies², and Thomas D. Tullius^{1,3}

¹Program in Bioinformatics, Boston University, Boston, Massachusetts, USA;

²National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA; ³Department of Chemistry, Boston University, Boston, Massachusetts, USA.

GC content has been shown to be an important aspect of human genomic function. Extending beyond the scope of GC content alone, there is a class of regions in the genome that have especially high GC content and are enriched for the CG dinucleotide - called CpG islands. CpG islands have often been linked to biologically functional genomic elements. DNA structure also contributes to biological function. Recent studies found that some DNA structural properties are correlated with CpG island functionality (Bock et al., 2006; Greenbaum, Parker, and Tullius, 2007). Here, we use hydroxyl radical cleavage patterns as a measure of DNA structure, to explore the relationship between GC content and fine-scale DNA structure. We show that there is a positive correlation between GC content and the solvent-accessible structural properties of a DNA sequence, and that the strength of this correlation decreases as genomic resolution increases. We demonstrate that regions of the genome that have highly solvent-accessible DNA structure tend to overlap functional genomic elements. Our results suggest that DNA structural properties that are encoded in the genome are important for biological function, and that the highly solvent-accessible nature of high GC content regions and CpG islands may account for some of their functional properties.

References

Bock, C. et al. **CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure.** *PLoS Genetics* 2, e26 (2006).

Greenbaum, J.A., Parker, S.C.J. & Tullius, T.D. **Detection of DNA structural motifs in functional genomic elements.** *Genome Research* 17, 940-6 (2007)

Secondary Metabolite Gene Cluster Analysis by the Cumulative GC and DNA Curvature Profile in *Aspergillus fumigatus* Genome

Jin Hwan Do, Satoru Miyano

Human Genome Center, Institute of Medical Science, University of Tokyo

The identification of gene cluster of fungal secondary metabolite is important for its biotechnological application as well as characterization of fungal secondary metabolism. Most of prediction methods for gene cluster of secondary metabolite severely depend on homology searches. However, homology-based approach has intrinsic limitation to unknown or novel gene cluster. This paper examines the GC and DNA curvature profile of 26 gene clusters of secondary metabolite in the *A. fumigatus* genome to find out potential conserved signatures for fungal secondary metabolite gene cluster. We found hint of the possibility of making a DNA region including signature genes such as polyketide synthase (PKS), nonribosomal peptide synthase (NRPS) and/or dimethylallyl tryptophan synthase (DMATS) to be real fungal secondary metabolite gene cluster. That is, given the DNA region with window-averaged DNA curvature values below 0.18 for at least 20kb around signature genes such as PKS, NRPS and DMATS, this genomic region has high probability of secondary metabolite gene cluster. Our result could be used for identification of gene cluster of secondary metabolite in other filamentous fungi, especially for that severely regulated by LaeA or other proteins with similar function to LaeA.

A Novel Strategy to Search Concerted Transcription Factor Activities Using Gene Expression Profile and Genomic Data

Yosuke Hatanaka, Masao Nagasaki, Rui Yamaguchi, Takeshi Obayashi, Kazuyuki Numata, Seiya Imoto, Teppei Shimamura, Kengo Kinoshita, Kenta Nakai, Satoru Miyano

Human Genome Center, Institute of Medical Science, University of Tokyo

We propose a novel strategy to search concerted Transcription Factor (TF) activities using correlation between gene expression profiles and genomic sequences. Assuming similar promoter structure induces similar transcriptional regulation, and thus induces similar expression profile, we compared correlation between genes with similar promoter structures and genes with similar expression profiles. Comprehensive TF binding site prediction for all human genes were conducted, with 19,096 promoter regions upstream of Transcription Starting Site (TSS) given from dbTSS. To optimize the TF binding site prediction, we filtered the predicted binding sites with co-expression genes data, provided from COXPRESSdb. Those promoter structure shared between co-expressed genes are finer prediction for TF binding sites, thus deduce finer prediction of concerted active TF groups.

Sensitivity-Driven Model Discrimination in the HOG Pathway in *S. Cerevisiae*

Christian Waltermann^{1,2}, Marcus Krantz², Stefan Hohmann², Edda Klipp¹

¹Computational Systems Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

²Institute for Cell and Molecular Biology, Gothenburg University, Gothenburg, Sweden

In previous experiments the high osmolarity glycerol (HOG) pathway elements Ssk1 and Pbs2 could be identified as highly sensitive towards over expression using the genetic tug-of-war (gTOW) method. Their severe gTOW phenotype could be rescued by deletion of HOG1, which encodes the MAPK of the HOG pathway, thereby linking over expression of Ssk1 and Pbs2 to growth-inhibitory constitutive activation of the pathway.

How precisely Ssk1 and Pbs2 are regulated however still remains to be elucidated. The current view of the pathway as well as existing mathematical models of the HOG MAPK cascade cannot account for the distribution of sensitivities observed in the gTOW experiment.

Here we present a number of mathematical models of how the sensitive nodes Ssk1 and Pbs2 might be regulated alternatively. Additional parameters are obtained using standard parameter estimation techniques to ensure that the output of the pathway remains in agreement with experimental data. The impact of different types of regulation on the sensitivity of the nodes is then discussed and proposed for further experimental testing.

How important is sequencing accuracy for next generation "deep" sequencing?

Jonas Maaskola

Progress in the development of new sequencing technologies mainly aims to increase data volume while at the same time sustaining sequencing accuracy. The typically desired sequencing error rate of below 1 in 1000 is historically founded on the necessity for very high accuracy in conventional, low-sequence-number cloning approaches. Given the redundancy of deep sequencing, such a high accuracy may be unnecessary for specific applications.

I studied the impact of lower sequencing accuracy on mapping precision and recall. The results suggest that, at least for some applications, the trade-off between read abundance, sequencing accuracy and costs could be resolved in favor of cheaper, "sloppier" machines with higher through-put.

Domain-Based Model for Protein Interaction Network

Morihiro Hayashida

Bioinformatics Center, Institute for Chemical Research, Kyoto University

Understanding protein functions and protein-protein interactions is essential to understanding biological processes. Many methods have been developed for inferring protein-protein interactions from protein sequence data. Among them, this talk focuses on methods based on domain-domain interactions, where a domain is defined as a region within a protein that either provides a specific function or composes a stable structural unit. In these methods, the probabilities of domain-domain interactions are inferred first from known protein-protein interaction data and protein domain data. Then, interactions between proteins are predicted using the probabilities of domain-domain interactions and the compositions of domains in the given proteins. This talk overviews several methods, which include the association methods, EM method, and LP-based methods. Moreover, this talk also reviews a simple evolutionary model of protein domains. Some kind of network based on the composition of domains shows two types of power-law behaviors. By combining it with a domain-based protein interaction model, a scale-free distribution of protein-protein interaction networks is obtained.

Understanding the Regulatory Mechanisms of Retrotransposed Genes in the *Drosophila* genus

Lan Hu, Daniel Segrè, Temple F. Smith

Graduate Program in Bioinformatics, Boston University, Boston, MA 02215

Retrotransposed genes are genes that are retrotranscribed from mRNA back into DNA, and inserted into a different location in the genome during the evolution. Thirty-four such genes have been reported in the *Drosophila* genus [1, 2]. The issue is how these genes are regulated or even actively used in which the original regulatory signals are either missing or incompatible with the new position. In this study, we analyzed a subset of the retrotransposed genes that are supported by multiple *Drosophila* species and their potential transcription factor binding sites. We propose two mechanisms that could regulate the retrotransposed genes at the new position. One possible mechanism, supported by the gene CG32119, suggests that the retrotransposed orthologs have been fortuitously inserted next to a gene that has compatible regulatory sites. A second possible mechanism, supported by the gene I(1)10Bb, suggests the co-existence of the original copy and the retrotransposed copy could help the regulatory elements to evolve for the retrotransposed copy, hence keep the function of the gene when the original copy was degenerated.

Reference

- [1] Bhutkar, A. et al. **Genome-scale analysis of positionally relocated genes**, *Genome Research*, 17(12), 2007.
- [2] **Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny**. *Nature* 450, 203-218, 2007

Modeling the Gene Expression of IL-2 in Single T Cells

Manuela Benary

Interleukin-2 (IL-2) is one of the first cytokines to be expressed by T helper cells (Th cells) after encountering an antigen. In contrast regulatory T cells do not produce IL-2, although they are similarly activated, except that an additional transcription factor (FoxP3) is expressed. The logic of gene expression of IL-2 is well understood (for review see [KIL06]). As more and more data for single cells based on fluorescent-activated cell sorting (FACS) (for example [PBZ+07]) become available, the gene expression of IL-2 in single cells can be described by appropriate mathematical models.

Based on statistical tests of available FACS experiments it seems reasonable that not only the amount of the transcription factors NFAT, NFkB and FOXP3 is important but also the ratio between the different transcription factors.

I will present an initial small model of the IL-2 gene expression on a single cell level. The model is adapted to data of regulatory T cells and T helper cells. I will discuss the results and arising problems when modelling gene expression on a single cell level.

Reference

- [PBZ+07] Podtschaske, M., U. Benary, S. Zwinger, T. Hofer, A. Radbruch and R. Baumgrass. **Digital NFATc2 activation per cell transforms graded T cell receptor activation into an all-or-none IL-2 expression.** *PLoS ONE*, 2(9):e935, 2007.
- [KIL06] Kim, H. P., J. Imbert and W. J. Leonard. **Both integrated and differential regulation of components of the IL-2/IL-2 receptor system.** *Cytokine Growth Factor Rev*, 17(5):349-66, 2006.

Cross-species analysis of gene expression

Marta Luksza

We present a method for genome-wide comparative cross-species analysis of expression data. A key component of our approach is the correct normalization of (logarithmic) expression levels, which results in a well-defined statistics of the data sets. In particular, we construct a quantitative mapping of samples (e.g., experimental conditions or tissues) both within and between species, which accounts for statistical redundancies and introduces a refined metric for expression vectors.

The similarity measure is then used in analysis of genome-wide human and mouse expression data. According to our measure, genes that have many co-expression interactions within one species tend to have more conserved expression patterns. We propose a statistically based algorithm for detection of genes that are significantly clustered in both species. The clusters we find suggest conservation of tissue specific expression and also of some specific pathways.

SuperToxic: a Systems Biology Approach to Toxicity

Swantje Struck, Ulrike Schmidt, Björn Grüning, Robert Preissner

Toxicity is by definition a measure of the degree to which something is able to produce illness or damage to an exposed organism (or part thereof). A central concept of toxicology is the dose-dependence, expressing that even water can lead to water-intoxication. On the other hand, there is a dose threshold, below which even the most toxic substance is nonhazardous. From this general point-of-view each and any compound should be included in a toxicity database and the particular assay (animal, tissue, cell etc.) the toxicity-value refers to, has to be defined. Furthermore, for an improved understanding of toxicity the molecular targets and the addressed pathways are of great importance.

For toxins, it is of special interest to consider the metabolic pathways, which lead to their production or degradation. Using inter-species comparisons, it will be possible, to understand the developments of the involved enzymes in the light of evolutionary pressure. Here, we undertake a first step towards the critical integration of toxicity data from different resources into a systems biology context. To this end, we compare the compounds to the KEGG ligands and map them onto the KEGG pathways (via targets and via biogenesis). In order to make a wide range of structural diversity available, we have collected toxicity information about approximately 50,000 compounds so far, which are stored in a database, called SuperToxic. Additional functionality like similarity or substructure searching, and links to associated pathways enable the identification of toxic substances with their corresponding sphere of activity. In pharmacological research this approach will be helpful to predict/avoid adverse effects of drugs.

Identification of Optimal Drug Targets in Differential Equation Networks

Marvin Schulz, Edda Klipp

Computational Systems Biology Group, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

During the stages of the development of a potent drug, candidate compounds can fail for several reasons. One of them, the efficacy of a candidate, can be tested *in silico* if an appropriate ordinary differential equation model of the affected pathway is available. With such a model at hand it is also possible to detect reactions having a large effect on a certain variable like a substance concentration.

We systematically test the influence of activators and inhibitors of different type and strength acting at different positions in the network. The effect on a quantity to be selected (e.g. a steady state flux or concentration) is calculated. Moreover, combinations are analyzed of two inhibitors or one inhibitor and one activator targeting different network positions.

Here, we present TIdc (Target Identification), an open source, platform independent tool to investigate ODE models in the common SBML format. It automatically assigns the respectively altered kinetics to the inhibited or activated reactions, performs the necessary calculations, and provides a graphical output of the analysis results. As an illustrative application, it is used to detect optimal inhibitor positions in simple branching networks and in a well studied model of glycolysis.

Network analysis of adverse drug interactions

**Masataka Takarabe¹, Shujiro Okuda¹, Masumi Itoh¹, Toshiaki Tokimatsu¹,
Susumu Goto¹ and Minoru Kanehisa^{1,2}**

¹ Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

² Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, Japan

Harmful effects associated with use of drugs are caused as a result of their side effects and combined use of different drugs. These drug interactions lead to increase or decrease of drug effects or other serious reactions. The known information about the potential risk of drug interactions are described in drug package inserts. Japanese drug package inserts are stored in the JAPIC (Japan Pharmaceutical Information Center) database and GenomeNet provides the GenomeNet pharmaceutical products database, which integrated the JAPIC and the KEGG database. In this study, we used the information on the drug package inserts and created drug-drug interaction networks to understand characteristics of drug interactions.

We extracted drug interaction data from the Japanese drug package insert stored in the GenomeNet pharmaceutical products database. Drugs or classes of drugs which cause adverse interactions with the drug are listed in each entry, and these interactions are classified according to risks, contraindications or cautions, for coadministration, and some entries include information about enzymes metabolizing the drugs. We defined drug target and drug-metabolizing enzyme as interaction factors using the information of them on KEGG DRUG. We created graphs of drug-drug interaction networks using the drug interaction data. In the resulting drug-drug interaction network, the drugs that associated with the same interaction factors were closely interconnected respectively.

Sampling Protein-Protein Geometries in Real Space

Aysam Gürler, Ernst-Walter Knapp

Freie Universität Berlin, Institut für Chemie und Biochemie
Takustr. 6, 14195, Berlin-Dahlem, Germany

Protein-protein docking is a major problem in structural biology. In general, the geometries of protein pairs are sampled by generating geometries, analyzing them with scoring functions and selecting appropriate structures for further refinement [1]. Here, we present a real space algorithm to sample geometries of protein pairs, which employs a simple scoring device. This sampling method generates structures of protein pairs as follows: (i) We reorient one protein of the protein pair using evenly distributed points on the surface of a sphere. (ii) The two proteins are translated along the line connecting surface points belonging to different proteins such that the surface points coincide. We applied this approach to a well established benchmark set of 22 enzyme-inhibitor complexes [2]. The resulting protein pair geometries were analyzed and selected using an amino acid and an atom pair based scoring function whose parameterization was essentially taken from the literature. The resulting ensemble of protein pair geometries illustrates that an efficient and robust sampling of near-native protein-protein decoys in real space is feasible. In this study, we have demonstrated that a discretisation of the rigid-body search in real space for protein-protein geometries provides an efficient and robust sampling scheme. In future work, we plan to employ this approach for well established scoring schemes and to introduce new scoring strategies.

Keywords: protein-protein docking, sampling protein pair geometries.

Fully Flexible Refinement of Docking Decoys

Stephan Lorenzen and Ernst-Walter Knapp

Since the experimental determination of structures of protein complexes is expensive and time consuming and often even impossible, computational protein--protein docking provides an interesting alternative to investigate the interactions of proteins with each other. By current Fast Fourier Transform (FFT) methods, billions of possible orientations of two proteins relative to each other can be generated and preselected efficiently using a simple scoring function based on shape and charge complementarity. However, a more detailed ranking and refinement of the docking decoys still poses a major problem. Here, we present a method to refine docking decoys using a replica exchange Monte Carlo (MC) approach where each MC step comprises a quick local minimization.

The replica exchange approach allows to cross energy barriers between different minima, and the resulting trajectory of decoys in the course of the MC simulation is an indicator of the width of the binding funnel.

After remodelling the side chains based on a rotamer library sampling the space of possible conformers, the minimization step provides full flexibility of both monomers, including the backbone.

Computer Aided Selection of Isotopomer Labels for Tracer Experiments

Benjamin Menküc and Christoph Gille

Radioactive tracer experiments are indispensable for the determination of flux rates in already known pathways as well as for the identification of new pathways. The information gained from such experiments depends on the labelling of the initial tracer metabolite, i.e. the atomic positions carrying a radioactive isotope. Here we present an algorithm and a software tool that facilitate the set up of an optimal carbon labelling pattern that assures the label to disseminate predominantly into those parts of the network under study. Our implementation is based on carbon fate maps and distinguishes between homotope and prochiral atoms. In addition, the software can be used to generate carbon transition probability matrices that can be used for the study of biochemical reaction mechanisms. In this article we present the algorithms and show an application of the software for the optimal estimation of flux rates in the intermediary metabolism of *E.coli*.

Documenting protein alignments

Christoph Gille, Andreas Hoppe, Hermann-Georg Holzhütter

Charite Universitätsmedizin Berlin, Institut für Biochemie
Monbijoustrasse 2, 10117, Berlin, Germany

Background

Reconstructed Complex biological networks are the essence of knowledge originating from experiments and modelling and gathered from scientific literature and databases.

Proteins are the major players in biological networks. Since knowledge on some individual proteins is still incomplete, their biological function is often deduced from similar homologous proteins which are already experimentally characterized.

As such indirect assignments are not as reliable as experimental evidences, they should be well documented and reviewed as soon as experimental data is available. Inconsistent operation of the resulting network may indicate invalid functional assignments.

Results

Here we present a novel feature of the alignment viewer STRAP which allows convenient publication of annotated sequence and 3D-structure alignments in form of a simple Web link to the Java-application STRAP. References to public file collections such as EMBL, KEGG, GeneBank, PDB, Pfam, Prodom, UniProt/SwissProt are encoded in the Web-link.

Data not included in public databases such as site specific annotations, 3D-rendering commands, yet unpublished sequence and 3D-data, 3D-transformation matrices must be stored on the Web-server in one additional file. In contrast, sequence features such as active site residues, phosphorylation sites and ligand binding sites usually do not need to be specified, as these data can be automatically retrieved and updated from public databases.

The alignment viewer may be of interest for many experimentalists, as it can be used to document sites of interest in a protein family under experimental investigation in project Web sites.

Availability

The STRAP program is published under the GNU-license condition and is Web-started from <http://3d-alignment.eu/>.

MicroRNA Diagnostic Panels and Gene Targets in ccRCC

Gyan Bhanot^{156*}, H. Liu¹, G. Alexe², D. Juan³, T. Antes³, C. Delisi⁴, L. Liou³
and S. Ganesan^{5^}

MicroRNAs (miRNAs) are a class of naturally occurring noncoding RNAs that regulate protein expression by targeting the mRNA of protein coding genes. The genes targeted by miRNA are involved in a wide range of biological functions, including cellular differentiation, development and apoptosis. Because of the large number of putative target genes that a given miRNAs can regulate, it is likely that they can perform both as oncogenes or tumor suppressor genes in a tissue or context specific manner. At present, computational predictions on putative targets of miRNAs are primarily based on searching for miRNA binding sites in the conserved 3'UTR region of genes using free energies of RNA-RNA duplexes[3][4][7]. However, because of the short length of miRNA sequences (18-24 nucleotides), these methods face a significant challenges in controlling the false positive rate. It is often difficult to find consistent predictions from different algorithms.

In this study, we propose and demonstrate a direct procedure to identify miRNA targets by a comparative analysis of the expression levels of matched miRNA and mRNA from the same patient cohort within a given phenotype (clear-cell Renal Cell Carcinoma or ccRCC cells vs normal kidney cells in the example considered here). Genes whose expression levels are highly correlated with the miRNAs that are differentially expressed in tumor tissue compared to normal tissue are identified as candidate targets. We find that a gene enrichment analysis of the differentially expressed miRNA/mRNA from renal cell carcinoma (RCC) compared to normal tissue identifies many known tumor markers and biological pathways regulated by miRNAs and suggests some new ones. The miRNA and mRNA identified by our analysis also provide excellent diagnostic panels for distinguishing normal kidney from ccRCC.

¹ BioMaPS Institute, Rutgers University, Piscataway, NJ 08854. Email: liuhq@biomaps.rutgers.edu.

² The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge MA, 02142

³ Boston University School of Medicine, 715 Albany St, Boston, MA 02118

⁴ Bioinformatics Program, 24 Cummington Street, Boston University, Boston, MA 02215

⁵ Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08093

⁶ Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540

Joint Senior Authors

Abstracts for the poster session

A Thermodynamic Approach for Modelling Cation Homeostasis in *Saccharomyces cerevisiae*

Susanne Gerber¹, Simon Borger¹, Svetlana Shabala², Sergey Shabala², Hella Lichtenberg-Fraté³, Jost Ludwig³, Mathias Kahm⁴, Maik Kschischo⁴, Edda Klipp¹

¹ MPI for Molecular Genetics, Computational Systems Biology Group, Germany

² University of Tasmania, School of Agricultural Science, Australia

³ University of Bonn, IZMB, Germany

⁴ University of Applied Science Koblenz; Germany

The focus of our modelling approach lies in introducing a systematic thermodynamically based description to explain mechanisms and interdependencies of selected plasma membrane cation transport systems in the yeast *S. cerevisiae*. This description employs the classical concept of linear flow-force relationships. The goal of this approach is to identify necessary and sufficient parameters and data sets for a comprehensive mechanistic description.

Relevant cations and their transport systems comprise potassium [K⁺], sodium [Na⁺], and protons [H⁺] and, to account for electro neutrality, the anion chloride [Cl⁻]. In addition, negatively charged polyanions of the cell such as proteins and nucleic acids will be taken into account. For a start, the computational model assumes changes of the intracellular cation concentrations as to be driven by the electrochemical gradients. Currently, each cation transport is represented by a single term. In the future, we intend to assign to these single terms detailed mechanisms of the individual transport systems. Upon external stimuli like e.g. increasing the extracellular potassium concentration, the combination of all equations may enable the prediction of changes of the internal ion concentrations as being mediated by the overall activity of involved transport systems. We intend to address the issue of dynamic integration with regard to the influence of pH[in] and pH[out], the membrane potential (Ψ) and the cell volume besides the driving force of electrochemical gradients (as basic start assumption). To account for the individual functional roles and transport mechanisms it is necessary to distinguish between (and differentially describe) ion channels, ion pumps and carriers. At any time, each of these systems is separately dependent on the membrane potential, the pH, ionic concentrations, the number of the membrane proteins (effects of gene regulation) and their activity (influenced by modifications like e.g. de-/phosphorylation).

The challenges of identifying the essential components for the achievement of a homeostatic concentration balance will be addressed as well as the limitations of the system. Necessary validations of the model based on available ionic flux data and so far known kinetic parameters will be discussed, also addressing further consecutive working steps.

Modeling DNA Replication in *Saccharomyces cerevisiae*

Thomas W. Spiesser, Edda Klipp, Matteo Barberis

Max Planck Institute for Molecular Genetics, Ihnestr 63-73, 14195 Berlin, Germany

In eukaryotes DNA replication is considered to proceed according to a precise program in which each chromosomal region is duplicated in a defined temporal order. However, recent studies reveal an intrinsic temporal disorder in the replication of yeast chromosome VI. Here we provide a model of the chromosomal duplication to study the temporal sequence of origin activation in budding yeast. The model is composed of four parameters that influence the DNA replication system: the lengths of the chromosomes, the explicit chromosomal positions for all replication origins as well as their distinct initiation times and the replication fork migration rate. The designed model is able to reproduce the available experimental data in form of replication profiles. The dynamic of DNA replication was monitored during simulations of normal and randomly perturbed replication conditions. Severe loss of origin function showed only little influence on the replication dynamics, wherefore systematic deletions of origins were simulated to provide predictions to be tested experimentally. The simulations provide new insights into the complex system of DNA replication, enabling us to explain trends exhibited during the DNA replication process and review the existing experimental data in the light of the above defined parameters.

Identification and Classification of Virulence Factors Using Phylogenetic Profiling

Roland Krause

The genomes of bacterial pathogens encode numerous genes for which no function has been found. Bioinformatic analysis like the use phylogenetic profiles have promised to generate hypotheses for functional traits but in the absence of large scale functional data, the establishment of these methods face many difficulties.

We explore how to use phylogenetic profiles for the detection of virulence factors of bacterial pathogens using curated profiles and functional screens for validation of the predictions. We find that several commonly used approaches perform poorly in these settings and propose a solution, which can also be applied to other problem instances.

Interaction of Signalling Pathways: cAMP and Calcium

Roberta Bianca Sprîncenatu

Max-Delbrück-Centre for Molecular Medicine

The two pathways, which contain cyclic adenosine monophosphate (cAMP) and inositol 1,4,5-trisphosphate (IP₃) as second messengers, are very important in signal transduction in biological cells. They have different effects, for example on muscle contraction, cellular motility, regulation of enzyme activity (IP₃ pathway) and on regulation of glycogen, sugar and lipid metabolism (cAMP pathway) depending on the cell type. The cAMP and IP₃ pathways have both been subject to extensive mathematical modelling, but so far they have been considered independently. However, experimental results suggest that they interact, such interactions being important for the effects they carry in cells. For a better understanding of the importance of these interactions we will develop a mathematical model reflecting the most important properties of this biological system in blowfly salivary gland *Caliphora vicina* and human embryonic kidney cells.

Comparative VEGF Receptor Tyrosine Kinase Modelling for the Development of Highly Specific Inhibitors of Tumor Angiogenesis

Ulrike Schmidt, Jessica Ahmed, Elke Michalsky, M. Hoepfner, Robert Preissner

The receptors of the Vascular Endothelial Growth Factor (VEGF-R) play a significant role in tumor development and angiogenesis and are therefore interesting targets in cancer therapy. The targeting of the VEGF-R is of special importance as the feed of the tumor has to be reduced. In general, this can be carried out by inhibiting the tyrosine kinase function of the VEGF-R. Nevertheless, there arise some problems with the specificity of known kinase inhibitors: they bind to the ATP-binding site and inhibit a number of kinases or the so far most specific inhibitors act at least on these three major types of VEGF-Rs: Flt-1, Flk-1/KDR, Flt-4. The goal would be a selective VEGF-R2 (Flk-1/KDR) inhibitor, because this receptor triggers rather unspecific signals from VEGF-A, -C, -D and -E.

Here, we describe the protocol starting from an established inhibitor (Vatalanib) with 2D-/3D-searching and property filtering of the in silico screening hits and the “negative docking approach”. In this way, we were able to identify a compound, which inhibits the VEGF-R2 with an IC₅₀ of < 5nM, while the VEGF-R1 is inhibited in the mM range. The outcome is not only a putative new anti-cancer drug but a confirmation of the holistic approach, considering not just the target but even the non-targets.

Keywords: in silico screening, homology modelling, docking, tyrosine kinase, angiogenesis, cancer, VEGF-R (vascular endothelial growth factor receptor)

Disorder as a Determinant of Evolution in Protein Microenvironments

Eric Franzosa, Yu Xia

Boston University Bioinformatics Program and Department of Chemistry

Current measures of whole-protein structural properties explain only a small fraction of evolutionary rate variation. This is surprising given that structure mediates most aspects of a protein's existence. We have investigated this paradox by looking for relationships between selective constraint and biophysical properties in protein microenvironments. We mapped genomic sequences from seven related yeast species onto roughly 500 homologous protein structures. Measured physical properties at residue sites (e.g., buried and interfacial surface area) could then be paired with the conservation status of the site across species. We uncovered strong functional relationships between measures of molecular disorder and residue mutation probability. These findings contribute to an improved understanding of the role that structure plays in protein evolution.

Conservation of Histone Marks and the PRC1 Component Ring1B at Promoters in Mouse and Human Embryonic Stem Cells

Esther Rheinbay^{1, 2}, Manching Ku^{2, 3}, Tarjei S. Mikkelsen², Eric Mendenhall^{2, 3}, Simon Kasif^{1, 4} and Bradley E. Bernstein^{2, 3, 5}

¹ Bioinformatics program, Boston University

² The Broad Institute of MIT and Harvard

³ Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital

⁴ Department of Biomedical Engineering, Boston University

⁵ Department of Pathology, Harvard Medical School

Recent development of chromatin-immunoprecipitation followed by ultra-high throughput sequencing has allowed efficient genome-wide mapping of various histone marks as well as chromatin-associated proteins. Histone modifications play a major role in the regulation of active and repressed chromatin through trithorax (TRX) and polycomb-repressive complexes (PRC), however the exact regulatory mechanisms remain unresolved.

We deployed conservation analysis to compare the activating histone mark H3K4me3 as well as the repressing mark H3K27me3 at 13,000+ promoters in human and mouse embryonic stem cells (ESC). Additionally, we compared occupancy of the polycomb-repressive complex 1 (PRC1) component Ring1B in both cell types.

We find that chromatin state as defined by histone modifications appears highly conserved between both species. The PRC1 component Ring1B occupies approximately half the sites marked by H3K27me3. Interestingly, most Ring1B-bound sites are promoters controlling developmental regulators. Moreover, H3K27me3 at promoters bound by PRC1 is more likely to be conserved in the other species. This is suggestive of active repression of PRC1-occupied promoters is a crucial memory mechanism during mammalian development.

Optimizing a Scoring Function for Protein-Protein Association

Florian Krull, Ernst-Walter Knapp

Institute of Chemistry and Biochemistry, Freie Universität Berlin

Multi-protein complexes are known to play a central role in many biological processes. A common way to predict the geometry of two interacting proteins is the usage of hierarchical two-step docking algorithms. Typically these algorithms generate a large set of possible protein complex geometries (decoys) in the first step, considering shape complementarity only. In the second step these decoys are re-ranked to discriminate near-native geometries from false positives. In our approach we use a scoring function based on residue pair contacts to score possible protein complexes. The knowledge-based weights of this scoring function are optimized by a method successfully applied in protein structure prediction. Furthermore, information of protein surfaces showing a high degree of residue conservation on the binding site is used.

Multiplierz: An Open-Source, Extensible Desktop Environment for Proteomics Data Analysis

Jignesh R. Parikh¹, Manor Askenazi^{2,3}, Nathaniel C. Blank², Tanya Cashorali², Yi Zhang², Scott B. Ficarro^{2,3}, Jarrod A. Marto^{2,3}

¹Bioinformatics Program, Boston University,

²Blais Proteomics Center and Department of Cancer Biology, Dana-Farber Cancer Institute,

³Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA

We present multiplierz, a novel and open-source software for mass spectrometry-based proteomics data analysis. Multiplierz acts as an integrated desktop environment by providing most tools required in a proteomics pipeline. A few proteomics integrated environments exist, most notably the Trans Proteomic Pipeline (TPP). However, multiplierz stands out in its extensibility and ease of use, making it ideal not only for integration into pipelines but also as a stand alone exploratory tool. We leverage existing commercial applications such as Microsoft Excel to generate image-enhanced spreadsheets that serve as information rich reports as well as easily alterable input sources. We have extended multiplierz to interact with OpenOffice.org Calc and are in the process of making multiplierz truly multi-platform software. We have used multiplierz to analyze data from various projects involving signaling downstream of oncogenic kinases, novel differentiation pathways in embryonic stem cells, and remodeling protein complexes. We demonstrate various features of multiplierz commonly used in such projects, including: (i) comparative analysis of protein abundance based on both stable isotope label and label-free approaches, (ii) functional annotation of identified proteins, and (iii) in-depth interrogation and confirmation of spectral features.

Prediction of optimal steady-state metabolic regulation using a linear constraint-based model

William Riehl and Daniel Segrè

Program in Bioinformatics, Boston University

Regulation of metabolic systems plays a crucial role in the function of cellular metabolism and the maintenance of a homeostatic steady state. This steady state is maintained through complex interactions among transcriptional and post-transcriptional mechanisms. While some of these mechanisms have been experimentally validated, many of the details of these interactions, including the various kinetic parameters, have proven difficult to elucidate. However, because it has been shown that evolution has driven metabolic networks to be optimal for several criteria, such as growth at one or more steady states, it is hypothesized that the regulatory system that operates on these networks has evolved to promote this optimal behavior. We further hypothesize that the topology of metabolic networks has evolved to be optimal for regulation that maintains the system around one or more steady states. Based on these hypotheses, we use methods related to flux balance analysis to construct a model of metabolic regulation based primarily on a metabolic network's stoichiometry, bypassing the requirement for the details of all kinetic parameters. This model predicts an optimal regulatory network of metabolic and genetic interactions that can resolve perturbations to a given steady state in a metabolic system. We use the model to predict optimal regulatory responses in both a simple toy network and in a fragment of *Escherichia coli* central carbon metabolism, and compare the results with published experimental data.

Determination of Time-Dependent Objectives in Metabolic Flux Models

Hsuan-Chao Chiu (hcchiu@bu.edu), Daniel Segrè (dsegre@bu.edu)

Graduate Program in Bioinformatics, Boston MA 02215, USA

Flux Balance Analysis (FBA) has been successfully applied to facilitate the understanding of cellular metabolism in model organisms. Current FBA research has mainly focused on studying the average performance of cell populations, without taking into account the metabolic state of each individual cell. However, both at the level of single cell and synchronized populations, the time-dependent dynamics could translate into time-dependent (e.g. cell cycle phase-specific) objectives and complex regulatory changes, whose understanding constitutes an open challenge. The knowledge of phase-specific objectives could help us better understand the principles of metabolic network operation during different growth phases in an organism's life cycle. We developed an FBA based approach to infer phase-specific objectives that are responsible for the determination of cellular fluxes in multiple time points. Specifically, our method infers the biomass compositions for different metabolic states based on a linear optimization framework that relies on given flux distributions. In addition, we analyze data in *E. coli* central carbon metabolism pathways to show the potential extension of our method to infer metabolic objectives based on gene expression data.

Improving Image Compression-based Algorithm for Measuring the Similarity of Protein Structures

Morihiro Hayashida¹, Tatsuya Akutsu¹(takutsu@kuicr.kyoto-u.ac.jp)

¹Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

1 Introduction

This poster proposes an improved method for measuring the similarity of protein structures. In the previous work [1], at first an original protein structure was transformed into a distance matrix, and the similarity of two protein structures was measured by a kind of compression ratio of the concatenated image. As image compression algorithms, JPEG, GIF, PNG, IFS, and SPC, were employed. The results of computational experiments suggested that SPC had the best performance. SPC is a lossless image compression algorithms developed by Said and Pearlman [2], which uses a simple pyramid multiresolution scheme enhanced with predictive coding. In this poster, the SPC algorithm is modified to compress two images without concatenating them to an image. We applied the modified method to clustering of protein structures, and obtained better results than that of SPC.

2 Similarity Measure

The similarity measure proposed in [1], AUSM, is derived from the universal similarity metric (USM) [3] based on Kolmogorov complexity. The Kolmogorov complexity $K(x)$ of an object x is defined to be the length of the shortest program P for a Universal Turing Machine U that outputs x . $K(x)$ is considered to be a measure of the amount of information contained in x , and can be approximated to be the compressed size $C(x)$ of x . Similarly, the conditional Kolmogorov complexity $K(x|y)$ is defined to be the length of the shortest program P to output x given y , and can be approximated to be $\max(C(x \cdot y), C(y \cdot x)) - C(y)$, where $x \cdot y$ denotes the concatenation of x and y . Then, AUSM is defined as follows:

$$AUSM(x, y) = \frac{\max(C(x \cdot y), C(y \cdot x)) - \min(C(x), C(y))}{\max(C(x), C(y))},$$

An object for a protein is a raw image given by PPM format constructed from the distance matrix $M_{ij} = \sqrt{(r^{(i)} - r^{(j)})^2}$, where $r^{(i)}$ denotes the three-dimensional coordinate of i th C-alpha atom.

The concatenation, $x \cdot y$, must be a rectangle image to be applied image compression algorithms. For that purpose, black regions were inserted to fill differences between two image sizes. In this poster, we propose a modified SPC method, MSPC, to deal with two images at once. SPC transforms a raw image by the method called S+P transform, and encode it by some kind of encoding algorithms. MSPC transforms two images by S+P transform, respectively, and encode them sequentially by an encoding algorithm.

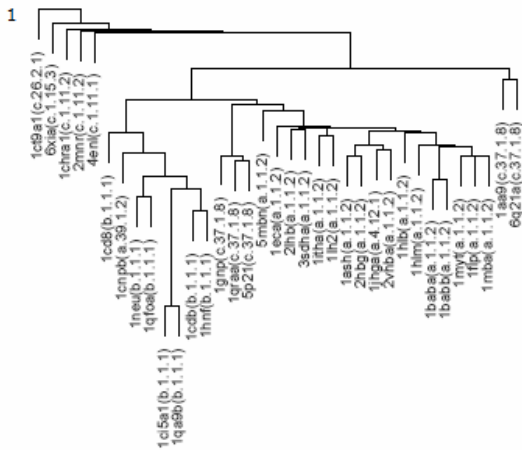


Figure 1: The clustering result using the nearest neighbor method for the MSPC method.

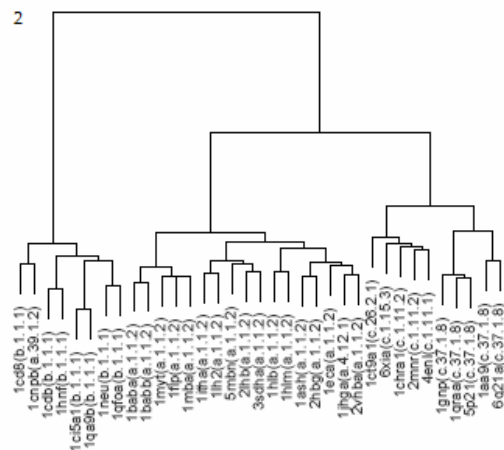


Figure 2: The clustering result using the Ward method for the MSPC method.

3 Computational Experiments

We used the same dataset in the previous work [1]. The dataset contains proteins identified by the following PDB codes: 18 all alpha proteins with Astral code “a” (1ash, 1eca, 1h1b, 1h1m, 1i1thA, 1mba, 1myt, 2hbg, 2lhb, 3sdhA, 1babA, 1babB, 1flp, 1lh2, 2vhbA, 5mbn, 1cnpB, 1jhga), 7 all beta proteins with Astral code “b” (1qa9B, 1cd8, 1cdb, 1ci5A, 1hnf, 1neu, 1qfoA), and 10 alpha and beta proteins with Astral code “c” (4enl, 2mnr, 1chrA1, 6xia, 1ct9A1, 1aa9, 1gnp, 1qraA, 5p21, 6q21A). Fig. 1 and 2 show the clustering results on the dataset for the MSPC method using the nearest neighbor and Ward method, respectively. As a result, MSPC, especially with the Ward method, classified the dataset better than SPC.

4 Discussion

In the previous result [1], alpha and beta proteins were classified mixed with all alpha proteins, whereas MSPC classified alpha and beta proteins separately from all alpha proteins correctly. Images are required to be rectangles even if they contain two images. We modified the SPC algorithm which had the best performance in the previous work, and calculated compressed sizes of two images without concatenating them to an image. In addition, values dealt with image compressions are restricted to integers of few bytes. In future work, we would like to develop compression algorithms for distances with real values.

Keywords: image compression, protein structure similarity

References

- [1] Hayashida, M. and Akutsu, T., *Proc. 6th Asia-Pacific Bioinformatics Conference*, 221–230, 2008.
- [2] Said, A. and Pearlman, W.A., *Proc. SPIE*, 2094:664–674, 1993.
- [3] Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P.E. and Zhang, H., *Bioinformatics*, 17:149– 154, 2001.

Canalisation as a plausible unifying mechanism for auxin transport in meristem development

Szymon Stoma

Shoot apical meristems continuously create new stems, leaves and flowers in highly precise positions. It is widely accepted, that the plant hormone auxin plays an important role in this process. This hormone is actively transported throughout the plant by proteins located at the plasma membranes of many cells. These transporters, so-called PIN-proteins, create fluxes of hormone in the plant that lead to the formation of local hormone maxima and minima which are subsequently interpreted in terms of differential cell behaviour. Two hypotheses have been used to explain the formation of the organised hormone fluxes. The first hypothesis proposes that auxin is transported against existing gradients. This concept can account for the observed auxin fluxes at the shoot apex and probably also for those in other tissues. The second hypothesis, called canalisation, proposes that the transporters act by amplifying and stabilising small existing fluxes. This concept is most widely used to explain the pattern of fluxes in internal tissues, but so far it was unclear whether it could also account for fluxes at the shoot apex. Here we demonstrate the results from the computational simulations of virtual tissues with the canalisation hypothesis.

These results led us to i) setting up an experimental framework that should allow us to test the validity of both proposed transport mechanisms ii) conclusion that the canalisation hypothesis is a plausible explanation of all observed auxin fluxes in the plant.

Refining a biomarker for lung cancer diagnosis that integrates gene expression and clinical variables

Adam C. Gower^{*2}, Jennifer Beane^{*1,2}, Paola Sebastiani³, Theodore H. Whitfield⁴, Katrina Steiling¹, Yves-Martine Dumas¹, Avrum Spira^{1,2}, Marc E. Lenburg^{1,2,5}

*Contributed equally to this work

¹The Pulmonary Center, Boston University Medical Center, Boston, MA ²Bioinformatics Program, Boston University, Boston, MA ³School of Public Health, Boston University, Boston, MA ⁴Biostatistics Solutions Consulting, Boston, MA ⁵Department of Genetics and Genomics, Boston University School of Medicine, Boston, MA

Lung cancer is the leading cause of cancer death, due in part to lack of early diagnosis. We have previously reported a gene expression profile in histologically normal large-airway epithelium obtained via bronchoscopic brushings that is an early diagnostic biomarker for lung cancer with greatly enhanced sensitivity relative to the cytology of materials obtained bronchoscopy. In the present work, our goal was to determine whether airway gene expression biomarkers capture information about lung cancer risk that is independent of clinical risk factors and to explore the potential benefits of a clinicogenomic model that combines gene expression with clinical risk factors.

Training (n = 76) and test sets (n = 52) were comprised of current and former smokers undergoing bronchoscopy for clinical suspicion of lung cancer at five medical centers. Data from these airway epithelial samples were used to develop and test logistic regression models that describe the risk of having lung cancer as a function of combinations of clinical risk factors with the multi-gene biomarker or the expression levels of individual genes, both of which we found to contain information about lung cancer risk that is independent of clinical risk factors. A clinicogenomic model constructed using the multi-gene biomarker and three clinical variables (age, mass size, and hilar or mediastinal lymphadenopathy) has improved performance relative to a model containing these clinical variables alone (sensitivity increases from 90% to 100%; specificity increases from 78% to 84% NPV increases from 93% to 100%; PPV increases from 72% to 79%).

Interestingly, we found that a clinicogenomic model containing these clinical factors and the expression levels of only six genes showed a similar performance benefit relative to the model containing the clinical variables alone. To determine whether the performance of the 6-gene clinicogenomic model is dependent on the predictive power of a specific set of genes, we iteratively removed the most predictive gene from the pool of available genes and built a new model. We found that many models with similar test set accuracy and specificity could be created, and are working to determine the utility of an ensemble of clinicogenomic models for making an ultimate diagnosis. The predictions of the clinicogenomic models were also compared with a subjective assessment of lung cancer risk made by pulmonary physicians blinded to cancer status in order to determine if the increased accuracy of the clinicogenomic model might be relevant for decisions about patient management. The clinicogenomic models had high accuracy where the physician assessments are most uncertain. Use of clinicogenomic models may expedite invasive testing and definitive therapy for smokers with lung cancer and reduce the number of invasive diagnostic procedures for individuals without lung cancer.

A stochastic model for p53-ERK dependent induction of apoptosis

Kazunari Mouri¹, Jose Nacher², Tatsuya Akutsu¹

¹ Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho Uji, Kyoto, 611-0011, Japan
² Future University Hakodate, Kamedanakano-cho Hakodate Hokkaido, 041-8655, Japan

The self protection mechanism inherent in cells against DNA damage and cellular stresses. In such mechanisms, cell-cycle arrest and repair of DNA damages exist for cellular survival. On the other hand, if the damages to a cell are excessive, the cell actively kills itself. This phenomenon is called apoptosis. Recently, the oxidative stress H_2O_2 can determine whether the cell should survive or induce apoptosis stochastically. In this study, our motivation is to clarify why the decision of the apoptosis is stochastic by constructing a network model of apoptosis at the molecular level. In previous experimental studies, it is suggested that the p53 tumor suppressor gene and the ERK protein are important for controlling apoptosis. The p53 gene expresses p53 proteins under cellular stress (we denote p53 as the gene and p53 as the protein product). After these proteins are phosphorylated for some time, the cell activates Caspase-3 and induces apoptosis. Meanwhile, ERK proteins suppress apoptosis by receiving survival signals. Recent experimental studies suggest that (i) the phosphorylation of ERK and p53 take place in separate cells, (ii) when the cellular stress (H_2O_2) increases, the number of cells which induce apoptosis gradually increases and the survival rate decreases. We seek to clarify such results by our model [4, 5].

At first H_2O_2 , which is one of the causes of DNA damage, harms a cell, and induces DNA double strand breaks (DSBs). In this situation, DNA repair proteins attach to the place where the DNA is damaged, and immediately begin to repair the DNA damage. Next, the complex of DNA and repair proteins activates ATM proteins, and they carry down the information of the existence of DNA damage [1]. Here, we assume that ATM* (activated ATM) activates p53, degrades Mdm2, and phosphorylates Mdm2. On the other hand, the stimulation of H_2O_2 activates Raf proteins, and signals phosphorylate ERK proteins [3]. Here, ERK* phosphorylates Mdm2, but ERK* is dephosphorylated by p53*. We find that these three proteins p53*, ERK*, and Mdm2* constitute positive feedback loop, which contributes to multi-stability of those proteins. In this multistability, when the p53* concentration is high, the ERK* concentration becomes low, and vice versa. Therefore, we can explain the experimental result (i) in the previous section by the positive feedback in p53*-ERK*-Mdm2* network.

Our model suggests that a positive feedback loop contributes to a binary decision of apoptosis and chemical reaction noise and variability of cells contribute to a stochastic decision of apoptosis under oxidative stress.

References

- [1] Bakkenist, C. J., and Kastan, M. B., *Nature*, 421:499–506, 2003.
- [2] Ma, L., Wagner, J., Rice, J. J., Hu, W., Levine, A. J., and Stolovitzky, G. A., *PNAS*, 102:14266–14271, 2005.
- [3] Malmföf, M., Roudier, E., Högberg, J., and Stenius, U., *J. Biol. Chem.*, 282:2288–2296, 2007.
- [4] Nair, V. D., Olanow, C. W., and Sealfon, S. C., *Biochem J.*, 373:25–32, 2003.
- [5] Nair, V. D., Yuen, T., Olanow, C. W., and Sealfon, S. C., *J Biol Chem*, 279:27494–27501, 2004.

Prediction of RNA Secondary Structure with Pseudoknots Using Integer Programming

Unyanee Poolsap, Yuki Kato and Tatsuya Akutsu

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan.

A molecule of RNA can be viewed as a single strand of the nucleotides (bases) adenine (A), guanine(G), cytosine (C) and uracil (U). A and U, C and G, and G and U can form a base pair via hydrogen bonding. Due to this property, an RNA strand can fold back on itself to form a secondary structure. We can represent a secondary structure of RNA by drawing a sequence of bases as a horizontal line and arcs above or below the sequence connecting two bases to represent the base pairs. If there are some crossing arcs, the secondary structure is said to contain pseudoknots. An often-used thermodynamic hypothesis states that RNA always forms a secondary structure with the lowest free energy. The problem of RNA secondary structure prediction is then modeled as an energy minimization problem, and many algorithms have been developed to solve it. Secondary structure without pseudoknot can be predicted in $O(n^3)$ time using dynamic programming algorithms (where n is length of the sequence) [2, 5]. However, it has been recognized that pseudoknots appear in many RNA molecules. Prediction of a secondary structure that contains pseudoknots is more difficult. Several existing algorithms can predict a secondary structure with pseudoknots in $O(n^4)$, $O(n^5)$ or $O(n^6)$ time. Moreover, prediction of an arbitrary planar secondary structure including pseudoknots is proved to be NP-hard [1]. We propose an integer programming (IP) based method to predict a pseudoknotted secondary structure of an RNA sequence. Despite the theoretical drawbacks in computational complexity, it is practical and reasonable to model the prediction of planar RNA pseudoknotted structure by IP formulation.

In our IP-based approach, we employ the stacking energy parameters for RNA folding at 37C as given in Mfold 3.0 [6]. We formulate a minimization IP model by defining the set of variables, the objective function and the set of constraints (see [3] for details). Then, we use the ILOG CPLEX 10.1, which is commercially available, to solve the IP model. The IP model was tested with the set of sequences with known structure. The sequences that are known to contain pseudoknots were obtained from PseudoBase. We selected 35 sequences in different families (viral 3UTR, mRNA, rRNA, ribozymes and tRNA-like) and of varying length (21-137 bases).

We evaluate the prediction results by calculating sensitivity, specificity and F-measure. We also compare our prediction accuracy with PKNOTS [4]. PKNOTS is an algorithm based on dynamic programming which can predict a pseudoknotted structure in $O(n^6)$ time. The average sensitivity of the IP-based method is slightly greater than that of PKNOTS. However, the specificity and F-measure are less than those of PKNOTS. If we look at the results of short sequences, the performance of IP-based method is comparable to PKNOTS.

References

- [1] Akutsu, T.: Discrete Applied Mathematics, Vol. 104, pp. 45–62 (2000).
- [2] Nussinov, R., Pieczenik, G., Griggs, J. R. and Kleitman, D. J.: SIAM Journal of Applied Mathematics, Vol. 35, No. 1, pp. 68–82 (1978).
- [3] Poolsap, U., Master's thesis, Kyoto University, Japan (2008).
- [4] Rivas, E. and Eddy, S. R.: Journal of Molecular Biology, Vol. 285, pp. 2053–2068 (1999).
- [5] Zuker, M. and Stiegler, P.: Nucleic Acids Research, Vol. 9, pp. 133–148 (1981).
- [6] <http://frontend.bioinfo.rpi.edu/zukerm/cgi-bin/efiles.cgi>

semanticSBML: A SBML user interface

Falko Krause

The System Biology Markup Language (SBML) is a common language for describing sets of biochemical reactions that are accompanied by kinetic information expressed by mathematical statements. The program semanticSBML allows users: to generate SBML documents from a list of KEGG reaction identifiers, to generate graph representations of models, to annotate models with MIRIAM annotations and to merge several models. The program includes an application programming interface (API), a console interface, and a graphical user interface (GUI).

Using Transcription Factor Binding Site Co-Occurrence to Predict Regulatory Regions

Holger Klein

Transcriptional regulation is known to be controlled by transcription factors (TFs) that form complexes with each other. Therefore binding sites for transcription factors (TFBSs) contained in these protein complexes often occur in proximity to each other. In this work we exploit this fact to identify TFs that are likely to interact with each other and to predict regulatory regions.

We annotate a set of non-redundant upstream regions of human genes with predicted transcription factor binding sites based on a representative set of vertebrate binding site motifs from the TRANSFAC database. We count co-occurring pairs of putative binding sites using a sliding window. Subsequently, significantly co-occurring pairs are identified using a log-odds score of observed and expected numbers of pairs of binding sites. To calculate the expected number of pairs we shuffle the TFBS labels and count pairs again repeatedly. Then we calculate the average number of co-occurrences. We assess the scoring procedure using known interactions of TFs from the TRANSFAC database. Scores for known combinations of TFs get significantly higher co-occurrence scores than combinations not known to interact. Furthermore we show pairs of TFs with high co-occurrence scores. Using predicted TFBSs and the co-occurrence scores shown above we construct TFBS-graphs, on which we calculate scores for the regulatory potential subsequently. The methods used include different ways of matching on the graph. We demonstrate the application of the methods on known regulatory regions. Comparison of maximum scores achieved for sets of known regulatory regions with the respective scores of random intergenic regions and artificial sets based on a permutation procedure shows the functioning of the procedures.

Ahmed, Jessica
Charité-Universitätsmedizin Berlin
Arnimalle 22, 14195 Berlin
jessica.ahmed@charite.de

Falcke, Martin
Max-Delbrück-Centre for Molecular Medicine
Robert- Rössle-Str. 10, 13125 Berlin
falcke@hmi.de

Bauer, Raphael
Charité-Universitätsmedizin Berlin
Arnimalle 22, 14195 Berlin
raphael.bauer@charite.de

Flöttmann, Max
Freie Universität Berlin
Innestr. 63-73, 14195 Berlin
floettma@molgen.mpg.de

Bausch, Johannes
Charité-Universitätsmedizin Berlin
Monbijoustr. 2, 10117 Berlin
johannes.bausch@charite.de

Franzosa, Eric
Boston University
24 Cummington St. Boston, MA 02215
franzosa@bu.edu

Basler, Georg
Universität Potsdam
Am Mühlenberg 1, 14476 Potsdam-Golm
basler@mpimp-golm.mpg.de

Gerber, Susanne
Max Planck Institute for Molecular Genetics
Boltzmannstr. 12, 14195 Berlin
gerber@molgen.mpg.de

Benary, Manuela
Humboldt-Universität zu Berlin
Invalidenstr. 42, 10115 Berlin
benaryma@cms.hu-berlin.de

Gille, Christoph
Charité-Universitätsmedizin Berlin
Monbijoustr. 2, 10117 Berlin
christoph.gille@charite.de

Bruck, József
Humboldt-Universität zu Berlin
Invalidenstr. 42, 10115 Berlin
j.bruck@freemail.hu

Gower, Adam
Boston University
715 Albany St. Boston, R-304, MA 02118
agower@bu.edu

Byrne David
Boston University
390 Commonwealth Ave Apt 501, Boston, MA
dbyrne@bu.edu

Grüning, Björn
Free University Berlin
Arminallee 22, 14195 Berlin
stefan.quenther@charite.de

Cui, Jike
Boston University
390 Commonwealth Ave Apt 501, Boston, MA
jike@bu.edu

Günther, Stefan
Charité Berlin
Arminallee 22, 14195 Berlin
stefan.quenther@charite.de

Chiu, Hsuan-Chao
Boston University
44 Cummington St. Boston, MA 02215
hchiu@bu.edu

Gürler, Aysam
Freie Universität Berlin
Fabeckstr. 36a, 14195 Berlin
querler@chemie.fu-berlin.de

Do, Jin Hwan
University of Tokyo, Human Genome Center
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639
jinhwan@ims.u-tokyo.ac.jp

Hancock, Timothy
Kyoto University, Bioinformatics Center
Gokasho, Uji, Kyoto 611-0011
timhancock@kuicr.kyoto-u.ac.jp

Ebenhöh, Oliver
Max Planck Institute for Molecular Plant
Am Mühlenberg 1, 14476 Potsdam-Golm
ebenhoeh@mpimp-golm.mpg.de

Handorf, Thomas
Humboldt-Universität zu Berlin
Invalidenstr. 42, 10115 Berlin
ThomasHandorf@gmx.de

Hatanaka, Yosuke
University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-86
hatanaka@hgc.jp

Hayashida, Morihiro
Kyoto University
Gokasho, Uji, Kyoto 611-0011
morihiro@kuicr.kyoto-u.ac.jp

Herzel, Hanspeter
Humboldt-Universität zu Berlin
Invalidenstr. 42, 10115 Berlin
herzelha@cms.hu-berlin.de

Holzhütter, Hermann-Georg
Charité-Universitätsmedizin Berlin
Monbijoustr. 2, 10117 Berlin
hergo@charite.de

Hu, Lan
Boston University
36 Cummington Street, Boston, MA 02215
hulan@bu.edu

Imoto, Seiya
University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639
imoto@imsu-tokyo.ac.jp

Jeong, Euna
University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639
eajeon@ims.u-tokyo.ac.jp

Karlstädt, Anja
Charite-Universitätsmedizin Berlin
Arnimallee 22, 14195 Berlin
anja.karlstaedt@charite.de

Klein, Holger
Max Planck Institute for Molecular Genetics
Innestr. 63-73, 14195 Berlin
holger.klein@molgen.mpg.de

Knapp, Ernst-Walter
Freie Universität Berlin
Takustr. 6, 14195 Berlin
knapp@chemie.fu-berlin.de

Kojima, Kaname
University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639
kaname@ims.u-tokyo.ac.jp

Krause, Falko
Max Planck Institut for Molecular Genetics
Innestr. 63-73, 14195 Berlin
krause_f@molgen.mpg.de

Krause, Roland
Max Planck Institute for Molecular Genetics
Innestr. 63-73, 14195 Berlin
roland.krause@molgen.mpg.de

Krull, Florian
Freie Universität Berlin
Fabeckstr. 36a, 14195 Berlin
fkroll@chemie.fu-berlin.de

Kruse, Kai
Humboldt-Universität zu Berlin
Invalidenstr. 42, 10115 Berlin
kaikruse@hu-berlin.de

Kühn, Clemens
Max Planck Institute for Molecular Genetics
Innestr. 63-73, 14195 Berlin
kuehn@molgen.mpg.de

Li, Xiaoqing
Max Planck Institute of Molecular Plant
Am Mühlenberg 1, 14476 Potsdam-Golm
xli@mpimp-golm.mpg.de

Lorenzen, Stephan
Freie Universität Berlin
Fabeckstr. 36a, 14195 Berlin
lorenzen@chemiefu-berlin.de

Luksza, Marta
Max Planck Institute for Molecular Genetics
Innestr. 63-73, 14195 Berlin
luksza@molgen.mpg.de

Mamitsuka, Hiroshi
Kyoto University
Gokasho, Uji, Kyoto 611-0011
mami@kuicr.kyoto-u.ac.jp

Maaskola, Jonas
Max-Delbrück-Centre for Molecular Medicine
Robert-Rössle-Str. 10, 13125 Berlin
jonas.maaskola@mdc-berlin.de

Menküc, Benjamin
Charité-Universitätsmedizin Berlin
Monbijoustr. 2, 10117 Berlin
benjamin@menkuec.de

Miyano, Satoru
University of Tokyo
4-6-1Shirokanedai, Minato-ku, Tokyo 108-8639
miyano@ims.u-tokyo.ac.jp

Schulz, Marvin
Max Planck Institute for Molecular Genetics
Innestr. 63-73, 14195 Berlin
schulzma@molgen.mpg.de

Mouri, Kazunari
Kyoto University
Gokasho, Uji, Kyoto 611-0011
kmouri@kuicr.kyoto-u.ac.jp

Shimizu, Yugo
Kyoto University
Gokasho, Uji, Kyoto 611-0011
shimizu@kuicr.kyoto-u.ac.jp

Numata, Jorge
Freie Universität Berlin
Fabeckstr. 36a, 14195 Berlin
numata@chemie.fu-berlin.de

Spiesser, Thomas W.
Max Planck Institute for Molecular Genetics
Innestr. 63-73, 14195 Berlin
spiesser@molgen.mpg.de

Parikh, Jignesh
Boston University
24 Cummington St. Boston, MA 02215
jig.parikh@gmail.com

Spr̄ncenatu, Roberta Bianca
Max-Delbrück-Centre for Molecular Medicine
Robert-Rössle-Str. 10, 13125 Berlin
bianca.sprincenatu@hmi.de

Parker, Steve
Boston University
44 Cummington St. Boston, MA 02215
parker@bu.edu

Snitkin, Evan
Boston University
44 Cummington St. Boston, MA 02215, USA
esnitkin@bu.edu

Poolsap, Unyanee
Kyoto University
Gokasho, Uji, Kyoto 611-0011
unyanee@kuicr.kyoto-u.ac.jp

Skupin, Alexander
Max-Delbrück-Center for Molecular Medicine
Robert-Rössle-Str. 10, 13125 Berlin
alexander.skupin@hmi.de

Preissner, Robert
Charité-Universitätsmedizin Berlin
Arnimallee 22, 14195 Berlin
robert.preissner@charite.de

Stoma, Szymon
INRIA
INRIA project-team virtual plants
szymon.stoma@inria.fr

Rajewsky, Nikolaus
Max-Delbrück-Centre for Molecular Medicine
Robert-Rössle-Str. 10, 13125 Berlin
rajewsky@mdc-berlin.de

Takarabe, Masataka
Kyoto University
Gokasho, Uji, Kyoto 611-0011
takarabe@kuicr.kyoto-u.ac.jp

Rheinbay, Esther
Boston University
390 Commonwealth Ave Apt 501, Boston, MA
rheinbay@bu.edu

Uhlendorf, Jannis
Max Planck Institute for Molecular Genetics
Innestr. 63-73, 14195 Berlin
uhlendorf@molgen.mpg.de

Riehl, William
Boston University
44 Cummington St. Boston, MA 02215
briehl@bu.edu

Waltermann, Christian
Max Planck Institute for Molecular Genetics
Innestr. 63-73, 14195 Berlin
christian.waltermann@molgen.mpg.de

Ross, Valery
University of Ottawa
550 Cumberland St. Ottawa, Ontario
vross047@uottawa.ca

Wan, Raymond
Kyoto University, Bioinformatics Centre
Gokasho, Uji, Kyoto 611-0011
rwan@kuicr.kyoto-u.ac.jp

Schuetz, Moritz
Max Planck Institute of Molecular Plant
Am Mühlenberg 1, 14476 Potsdam-Golm
schuetz@mpimp-golm.mpg.de

Wang, Connie
Yale University
cwang@chemie.fu-berlin.de

The Spreewald

The Landscape

About 100 kilometres from the German capital city Berlin, the Spreewald region stretches over a total area of more than three thousand (3,173) square kilometres, split into the Lower and Upper Spreewald. In its core, 75 km long and 15 km wide, is the area designated by UNESCO in 1991 as the Spreewald biosphere nature reserve.

This landscape of lowland and water meadows, unique in Central Europe, developed around 20,000 years ago in the Ice Age. The landscape was named after the Spree, a river with a 970 km long network of waterways winding through extensive meadows, forests and scattered settlements.

Several thousands of animal and plant species, many of them very rare and on the list of endangered species, can be found in the Spreewald. Hundreds of butterfly species, numerous dragonflies, mussels, snails and slugs, batrachians, fish, mammals, and birds have their home here. For example, the black and white stork, crane, white-tailed eagle, otter, hoopoe, or the curlew all live in the various biotopes.

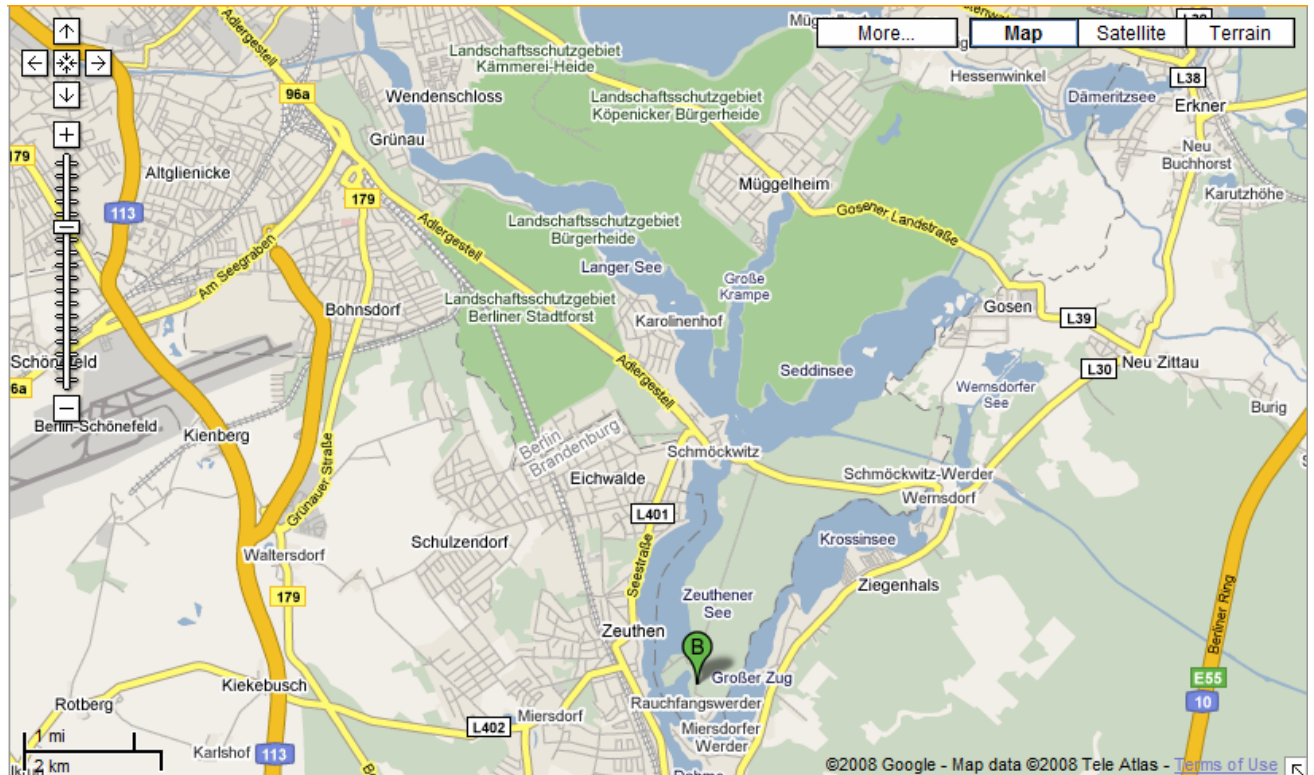
The flora of the Spreewald displays a similar variety. The banks of the small rivers are shadowed by slender alders, mighty poplars and oak-trees, creating a romantic leafy canopy above the streams. Various herbs such as Pfeilkraut, Wasserstern as well as water lilies are part of the rich aquatic plant world.

The development of the Spreewald as a cultivated landscape is closely connected to the people who have utilized and developed it for centuries. The very intense cultivation of large areas has unfortunately strongly affected the balance of nature. However efforts have been made over the last few years to repair the damage. So the Spreewald has been divided into different protected zones, whose core is kept untouched by humans. A number of individual measures help to maintain the natural habitats of plants and animals, but also ensure that other areas of the Spreewald are still available for economic use.

The Legend of the Spreewald's Origin

One legend tells that the Spreewald was formed as follows: The devil was ploughing fields in the region. He ploughed with two big black oxen, roaring and shouting terribly. The ploughing was hard and heavy going and the oxen were making a very slow job of it. So the devil took his hat and threw it, cursing, at the oxen. They got frightened, ran off and leapt criss-cross in all directions, still pulling the plough behind them. The deep furrows left by the plough caused many small streams to form where before there was one straight river bed.

Teikyo Hotel – map:



How to get there:

- Train S8 to S-Bahn station Grünau, then tram number 68 from Grünau to Alt-Schmöckwitz and finally bus number 168 from Alt-Schmöckwitz to Teikyo Hotel, where the Workshop takes place.

More information about the public transport in Berlin with journey planner and schedules:

- <http://www.bvg.de/index.php/en/Bvg/Start>