GCB 2012 Poster Abstracts

Table of Contents

Large-Scale Evolutionary Patterns of Protein Domain Distributions in Eukaryotes

Arli A. Parikesit, Peter F. Stadler, and Sonja J. Prohaska
Institute of Computer Science, University of Leipzig
arli@bioinf.uni-leipzig.de

The genomic inventory of protein domains is an important indicator of an organism's regulatory and metabolic capabilities. Existing gene annotations, however, can be plagued by substantial ascertainment biases that make it difficult to obtain and compare quantitative domain data [PSP10]. We find that quantitative trends across the Eukarya can be investigated based on a combination of gene prediction and standard domain annotation pipelines. Species-specific training is required, however, to account for the genomic peculiarities in many lineages [PSP11]. In contrast to earlier studies we find wide-spread statistically significant avoidance of protein domains associated with distinct functional high-level gene-ontology terms.

References

[PSP10] Arli A. Parikesit, Peter F. Stadler, and Sonja J. Prohaska. Quantitative Comparison of Genomic-Wide Protein Domain Distributions. In D. Schomburg and A. Grote, editors, *German Conference on Bioinformatics 2010*, volume P-173 of *Lecture Notes in Informatics*, pages 93–102, Bonn, 2010. Gesellschaft für Informatik.

[PSP11] Arli A. Parikesit, Peter F. Stadler, and J. Prohaska, Sonja. Evolution and Quantitative Comparison of Genome-Wide Protein Domain Distributions. *Genes*, 2:912–924, 2011.

e!DAL: A Framework for Storing, Sharing and Citing Primary Life Science Data

Daniel Arend, Matthias Lange, Christian Colmsee, Steffen Flemming,
Jinbo Chen and Uwe Scholz
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
D-06466 Stadt Seeland, OT Gatersleben, Germany
arendd@ipk-gatersleben.de

High troughput technologies produce a huge amount of primary data. In classic scientific publication process, primary data is usually aggregated to a number of paragraphs in a journal article and proven by figures, tables and supplementary material. So it is that the value of primary data, on which the scientific conclusions are based on, get increased attention in public as well as in the research community. This leads to novel strategies and concepts for primary data citation, which must be substantively underpinned by enhancements to classic data management systems.

Here we present the JAVA-based *e!DAL*-API, a comprehensive storage backend for primary data management. It stands for (**E**LECTRONICAL **D**ATA **A**RCHIVE **L**IBRARY) and implement a primary data storage infrastructure, but with an intuitive usability like a classical file system. Main features are *version and meta data management*, *data citations*, support for *information retrieval*, *persistent identifiers* and its *easy and modular integration* into existing data frontends and information systems. The API has been designed and tested using experiences from several research projects and literature studies.

Primary data preservation in life sciences is accompanied by enhanced requirements to data management systems. *e!DAL* combines this novel araising requirements with features known from file systems, databases, content management and version control to one homogeneous storage system. It supports an embedded use in stand-alone JAVA software or in server mode a data repositories for collaborative, remote accessible data services. Thus, *e!DAL* is an efficient complement for data frontends, information systems and data management systems. The JAVA libraries, Maven artifacts, sample code, a show case demo, and the API-documentation can be downloaded from:
`http://projects.ipk-gatersleben.de/eDAL-Project`

Involvement of mTor-pathway in aging during cultivation of primary mouse hepatocytes

Wolfgang Schmidt-Heck, Uwe Menzel and Reinhard Guthke
Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute (HKI), 07745 Jena, Germany
wolfgang.schmidt-heck@hki-jena.de

Liver is the main organ of intermediary metabolism. Lipids, amino acids and sugars are metabolized by hepatocytes according to the need of the body. Isolated hepatocytes can be used to study these topics of liver metabolism in vitro. Therefore, there is a need to understand signal transduction and metabolic pathways of hepatocytes. On the other hand, it is necessary to characterize the influence of the in vitro culture conditions on signal transduction and metabolism of hepatocytes.

Data of cultivated hepatocytes [ZEL10] was used to investigate the cellular response to cultural adaptation. To monitor changes at the transcription level, Affymetrix GeneChip MOE 430 2.0 oligonucleotide arrays were used for hybridisation. Six samples were taken within a period ranging from 3 to 48 hours after isolation. The mRNA of freshly harvested cells (3h = immediately after attachment) was used as control. The data were pre-processed using Bioconductor Software. 3362 probesets were found to be differentially expressed by a fold change greater three in one or more samples after isolation. Scaled expression profiles of the differentially expressed genes were clustered using Fuzzy c-means algorithm into eight groups. The result of transcriptome data analysis was mapped to the signalling pathways linking mTORC1 and mTORC2 to ageing via protein synthesis and autophagy [HAN09]. The activated sub-networks were described by a system of linear differential equations. The simulation of the obtained network shows an activation of the mTor pathway by nutrients activated signals (PI3K/Akt, Erk12/Rsk and p38/Mk2 signalling). The direct activation of „mTor complex 1“ through Nrf2 signalling was observed.

References

[ZEL10] Zellmer S, et al.: Transcription factors ETF, E2F, and SP-1 are involved in cytokine-independent proliferation of murine hepatocytes. Hepatology 2010, 52(6):2127–2136

[HAN09] Hands SL, et al.: mTOR's role in ageing: protein synthesis or autophagy? Aging. 2009:586–597.

Analysis of the relation of replication mechanism and nucleotide composition in metazoan mitogenomes

Abdullah H. Sahyoun*[1,2], Matthias Bernt[1], Peter F. Stadler[1], and Kifah Tout[2]

[1] *Institute of Bioinformatics, Leipzig University, Germany*
[2] *AZM Center for Biotechnology Research, Lebanese University, Lebanon*

Strand asymmetry of the nucleotide composition is a well known feature of animal mitochondrial genomes. There are differences in the composition along a single genome as well as when comparing genomes of different species. The understanding of the mutation processes modifying strand asymmetry is essential for a better understanding of the characteristics and evolution of mitochondrial genomes. Previous studies indicated that the differences in the nucleotide frequencies are associated with replication alone or both replication and transcription [FM97]. A well accepted hypothesis is that the replication process renders the heavy strand in a single stranded state for different amounts of time exposing it to an asymmetric mutation process [DA92, RAC98]. We aim to find the origins of replication of mitogenomes using the duration of single strandness of the heavy strand(D_{ssH}) and the asymmetry of the nucleotide composition as measured by AT- and GC-*skew*. To this end a linear regression approach is employed assuming a linear relation of the D_{ssH} and the dinucleotide-skews. Initial results are presented for several different animal clades.

References

[DA92] Clayton DA. Transcription and replication of animal mitochondrial DNA. *International Review of Cytology*, 141:217–232, 1992.

[FM97] Ochman H Francino MP. Strand asymmetries in DNA evolution. *Trends in Genetics*, 13:240–245, 1997.

[RAC98] Pesole G Reyes A, Gissi C and Saccone C. Asymmetrical Directional Mutation Pressure in the Mitochondrial Genome of Mammals. *Molecular Biology and Evolution*, 15:957–966, 1998.

*abdullah@bioinf.uni-leipzig.de

An Automaton-Based View on Error-Tolerant Pattern Matching with Backward Search

Dominik Kopczynski[†] and Sven Rahmann[†‡]
[†] *Collaborative Research Center (Sonderforschungsbereich, SFB) 876, Computer Science XI, TU Dortmund, Germany*
Dominik.Kopczynski@tu-dortmund.de
[‡] *Genome Informatics, Institute of Human Genetics, Faculty of Medicine, University of Duisburg-Essen, Germany*
Sven.Rahmann@uni-due.de

Backward search was introduced by Ferragina and Manzini and has become a standard index-based linear-time low-memory exact pattern search technique [FM00]. It is used as a computational core in many read mapping applications in the context of next generation sequencing data analysis. For this purpose, the backward search core is generally augmented by error tolerance, which complicates book-keeping and lends itself to error-prone implementations. Here we introduce an automaton-based view on error-tolerant backward search by combining the non-deterministic finite automaton from the error-tolerant Shift-And algorithm with exact backward search. This leads to a conceptually simple, efficient, easily implementable version of error-tolerant backward search that additionally is well suited for teaching in class.

Acknowledgments. The authors are supported by the Collaborative Research Center (Sonderforschungsbereich, SFB) 876 "Providing Information by Resource-Constrained Data Analysis" within project TB1 (http://sfb876.tu-dortmund.de).

References

[FM00] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE, 2000.

Modelling of Influenza Virus Infection in Mice

Himanshu Manchanda[1,2], Nora Seidel[2], Michaela Schmidtke[2], Reinhard Guthke[1]

[1]*Leibniz Institute for Natural Product Research and Infection Biology –Hans Knöll Institute, Jena*

[2]*Jena University Hospital, Institute for Virology and Antiviral Therapy, Jena*

Himanshu.manchanda@hki-jena.de

Influenza infections are of major medical and economical importance due to the high illness rate especially in infants, older people, and those with chronic diseases. The latter are at a high risk to develop severe disease with frequent hospitalisation and worldwide half a million deaths per year. For establishment and characterization of murine models for antiviral studies, there is a special need to study and quantify the dynamics of viral infection without and with drug treatment. Mathematical modelling have been proven to provide helpful tools in the analysis of viral infections.

To evaluate the course of influenza in mice , we started with the development and analysis of a simple Ordinary Differential Equations (ODE)-based mathematical model that consists of two time dependent variables and differential equations. One of them represents the kinetics of virus-induced disease reflecting the pathogenicity that we fitted to the measured disease activity score. The other variable represents the strength of the host's immune defence which dynamics we identified indirectly by the aforementioned fitting.

This model exhibits a rich range of solutions for various viral infections. We were able to fit the model to eleven data sets for four different viral strains each inoculated at 2 or 3 different infection doses. The model is simulated and fitted using the R packed FME [Soe10]. With extension of the basic model, we were able to fit the biphasic disease kinetics. This will help us in quantifying the pandemic or the resistant viral strains and the therapeutic efficiency of different drugs.

References

[Pug08] Andrea Pugliese and Alberto Gandolfi. A simple model of pathogen-immune dynamics including specific and non-specific immunity. Math Biosci. 214:73-80, 2008.

[Soe10] Karline Soetaert and Thomas Petzoldt. Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package FME. Journal of

Statistical Software, 33(3), 1-28, 2010.

A novel approach for genome-wide prediction of secondary metabolite gene clusters

Thomas Wolf, Vladimir Shelest, Ekaterina Shelest
Leibniz Institute for Natural Product Research and Infection Biology e. V. Hans-Knöll-Institute (HKI)
Research Group Systems Biology / Bioinformatics
thomas.wolf@hki-jena.de, ekaterina.shelest@hki-jena.de

Secondary metabolites (SMs) are pharmaceutically important natural products mostly produced by fungi and bacteria. Genes involved in the SM biosynthesis are often co-regulated and organized in clusters. Those can be regulated by cluster-specific transcription factors (TFs). The field of cluster prediction is quite important to secondary metabolite research, while the accuracy of currently available cluster prediction tools is less adequate.

We suggest a novel not-similarity based method to predict SM gene clusters, comprising the density of transcription factor binding site (TFBS) motifs. The occurrences of cluster-specific TFBSs should be higher in the cluster and less probable in other parts of the genome. Though, their occurrence outside the cluster is not excluded. Our algorithm searches for motif-enriched consecutive promoter regions. Gaps within the cluster are allowed. Initially, over-represented motifs in an interim set of promoters around the SM backbone gene are predicted. Subsequently, each significant motif is searched in all promoter sequences of the genome. The succession of promoters is scanned by a sliding window, counting the number of motifs for each frame. The highest motif density should be obtained for a frame equal to the cluster promoters, considering the possibility of different frame lengths.

Effectiveness of the method was successfully demonstrated by re-identifying several known clusters with wet-lab proven borders. We also show the applicability to completely unknown clusters.

Analysis of the transcriptome of rat liver in response to PPAR angonists

Martin Bens[1], Sebastian Vlaic[1], Reinhard Guthke[1] and Jürgen Borlak[2]

[1]*Research Group Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany.* [2]*Medical School of Hannover, Center for Pharmacology and Toxicology, Hannover, Germany.*
{martin.bens, sebastian.vlaic, reinhard.guthke}@hki-jena.de,
borlak.juergen@mh-hannover.de.de

A number of essential reactions in metabolism take place in the liver. Among others, these reactions are involved in detoxification of toxic compounds and lipid metabolism. Liver damage can lead to disregulation of these reactions and result in disorder of energy homeostasis. Peroxisome proliferator-activated receptors (PPARs) play an important role in modulating lipid metabolism and glucose homeostasis and are therefore an interesting therapeutic target useful in treatment of these misregulations. We investigated gene expression in rat liver after perturbation with PPAR-agonists clofibrate (CF) and diethylhexylphthalat (DEHP). The gene expression of 2 perturbated groups was measured and compared with the expression of a control group. Microarrays were used to detect the gene expression after 3, 7 and 28 days duration of treatment. Analysis of Gene Ontology and KEGG-Pathways assignments of the differentially expressed genes (DEGs) showed that both treatments increase the oxidation of fatty acids, though only the CF-group showed a significant increase of Pparα expression. In contrast to the overexpression of Pparα, the expression of Pparδ showed a tendency to reduction in both groups, however, it was significant only in the CF-group. Furthermore, a significant overexpression of Srebf1, an important transcription factor that regulates lipogenesis, was observed in the CF-group, which correlates with the increasing expression of Pparα over time. Additionally, CF regulates a wide range of genes which are involved in the cell cycle. The reduced expression of oncogenes like Myc or p21 suggests a reduction of tumor progression under the CF treatment. DEHP has only a minor influence on the regulation of cell cycle genes. Moreover, the number of DEGs in case of DEHP treatment was lower than under CF treatment (42 %).

MicroRNA-Seq data analysis for age-related comparison of mouse and short-lived fish Nothobranchius furzeri

A. Dix[1,3], S. Priebe[1,3], R. Guthke[1,3], M. Baumgart[2,3], A. Cellerino[2,3]
[1]*Leibniz Institute for Natural Product Research and Infection Biology e.V. - Hans-Knöll-Institute*
[2]*Leibniz Institute for Age Research - Fritz Lipmann Institute*
[3]*Jena Centre for Systems Biology of Ageing*
andreas.dix@hki-jena.de

Although many biological principles of the aging process remain to be elucidated, it is known that aging is an accumulation of changes and damage over time [LCM+05, BA04]. This damage results in or is partially caused by changes of gene expression and gene regulation. Since microRNAs (miRNAs) regulate about 30% of all animal genes, they are making a meaningful contribution to gene regulation [BC07]. Therefore, the analysis of miRNAs is of high importance for age research.

This work compares the influence of miRNAs on age-related genes between mouse and the turquoise killifish, *Nothobranchius furzeri*. This fish features an extremely short lifespan for a vertebrate making it a suitable organism for aging reasearch [BGP+12].

Throughout this study, numerous mouse miRNAs were identified as age-related. For many of them, the association to aging could be confirmed by other studies. Hence, there is a high probability that the other miRNAs are relevant for aging, too. In contrast to this, the number of age-related killifish miRNAs is much smaller and none of them could be verified as associated to aging by literature.

Additionally, conserved patterns of expression change were found for several miRNAs. In mouse, some of them are predicted to control the expression of age-related genes. However, an association to aging could not made for the conserved killifish miRNAs.

References

[BA04] R. L. Bowen and C. S. Atwood. Living and dying for sex. *Gerontology*, 50(5):265–290, 2004.

[BC07] N. Bushati and S. M. Cohen. MicroRNA functions. *Annual Review of Cell and Developmental Biology*, 23:175–205, 2007.

[BGP+12] M. Baumgart, M. Groth, S. Priebe, J. Appelt, R. Guthke, M. Platzer, and A. Cellerino. Age-dependent regulation of tumor-related microRNAs in the brain of the annual fish *Nothobranchius furzeri*. *Mechanisms of Ageing and Development*, 133(5):226–233, 2012.

[LCM+05] D. B. Lombard, K. F. Chua, R. Mostoslavsky, S. Franco, M. Gostissa, and F. W. Alt. DNA repair, genome stability, and aging. *Cell*, 120(4):497–512, 2005.

Structural Features of Protein-Protein Interfaces analyzed with Concepts of Information Theory

Christophe Jardin[1], Arno Stefani[2], Olaf Othersen[1], Johannes Huber[2] and Heinrich Sticht[1]

1) Institute for Biochemistry, FAU Erlangen-Nuremberg , Germany

2) Institute for Information Transmission, FAU Erlangen-Nuremberg, Germany

{Christophe.Jardin, Olaf.Othersen, h.sticht}@biochem.uni-erlangen.de

{stefani, jbhuber}@nt.e-technik.uni-erlangen.de

Molecular docking represents a versatile and important computational method for determining the structure of protein-protein complexes. Despite considerable efforts, a general solution to this problem is not yet within reach. One major challenge is the definition of suitable criteria for a scoring function that allows the identification of a good docking solution among many false arrangements.

Our previous work has demonstrated that the concepts from information theory can actually be adapted to treat the biological problem of protein-protein docking: a formalism has been developed, based on the concept of mutual information (MI), to investigate several structural features of the protein-protein docking solutions for their information content. We have also shown that the MI-values can successfully be converted into a scoring function [Oth11]. However, these first “proof-of-concepts” also emphasized aspects that had to be improved to result in a robust and widely applicable approach.

We present here an extended MI-based approach that relies on a larger dataset and allows a more flexible treatment of the structural features in the scoring function. The new training consists of carefully chosen docking solutions generated with the docking program FTDock. The role of amino acid diversity was investigated by comparing the information content of the different structural features when using different hierarchy of amino acid alphabets. A further improvement is the detection of redundancies between different features and the development of a suitable formalism for the estimation of the MI.

References

[Oth11] Olaf Othersen et al. *J. Mol. Model.*, 18(4):1285-1297, 2012.

MetaCrop: A Database for Plant Metabolism and Plant Metabolic Modelling

Stephan Weise[1], Christian Colmsee[1], Tobias Czauderna[1], Eva Grafahrend-Belau[1], Anja Hartmann[1], Astrid Junker[1], Björn H. Junker[1], Matthias Klapperstück[1], Uwe Scholz[1] and Falk Schreiber[1,2]
[1]*Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany and* [2]*Martin Luther University Halle-Wittenberg, Institute of Computer Science, Halle, Germany*

weise@ipk-gatersleben.de

Crop plants are of vital significance as a source of food, feed, energy and feedstock for the chemical industry. Given the close connection between plant metabolism and usability of plant products, there is a growing interest to understand and predict the behaviour and regulation of plant metabolic processes.

MetaCrop [SCC+12], now available in version 2.0, is a manually curated metabolic pathway database focussing on the primary metabolism of crop plants. MetaCrop contains in-depth information on the compound, reaction and pathway level including pathway maps, spatial/temporal location information, reaction kinetics and literature references. Intended to support crop plant research, MetaCrop allows researchers (i) to explore metabolic information by browsing through various levels of abstraction, (ii) to integrate experimental data into metabolic pathways and (iii) to create metabolic models for simulation purposes. MetaCrop is accessible via a web application (http://metacrop.ipk-gatersleben.de), web services and an add-on to the visualisation software Vanted.

References

[SCC+12] F. Schreiber, C. Colmsee, T. Czauderna, E. Grafahrend-Belau, A. Hartmann, A. Junker, B. H. Junker, M. Klapperstück, U. Scholz, and S. Weise. MetaCrop 2.0: managing and exploring information about crop plant metabolism. *Nucleic Acids Research*, 40(Database issue):D1173–D1177, 2012.

The mutually exclusive spliced exome of *Drosophila melanogaster*

Klas Hatje and Martin Kollmar
Max Planck Institute for Biophysical Chemistry, Göttingen, Germany
hakl@nmr.mpibpc.mpg.de

Alternative splicing is an important process in higher eukaryotes that allows generating several transcripts out of one gene. One type of alternative splicing is mutually exclusive splicing, which refers to the splicing of exactly one exon out of a cluster of neighbouring exons into the mature transcript. Recently, we introduced a new method to predict mutually exclusive spliced exons based on several preconditions to create biological meaningful transcripts [PH11]. These preconditions have been implemented into an extension to the gene structure reconstruction tool WebScipio (www.webscipio.org) [HK11]. The new algorithm was used to reconstruct the mutually exclusive exome of the model organism *Drosophila melanogaster* based on the advanced annotation of Flybase (release 5.36). Of the 259 mutually exclusive spliced exons already annotated in the *Drosophila melanogaster* genome, our prediction algorithm reconstructed 222, resulting in a sensitivity of 85.7 %. In total more than 600 mutually exclusive spliced exons were predicted with the same parameters, resulting in many new exon candidates, that are not annotated at all. To find further evidence for the predictions, the conservation of the mutually exclusive exons in 11 additional *Drosophila* species was analysed and expressed sequence tag (EST) data were mapped onto candidate genes. The data of all mutually exclusive exomes reconstructed so far are accessible via www.motorprotein.de/kassiopeia.

References

[HK11] Klas Hatje, Oliver Keller, Björn Hammesfahr, Holger Pillmann Stephan Waack and Martin Kollmar. Cross-species protein sequence and gene structure prediction with fine-tuned Webscipio 2.0 and Scipio. *BMC Research Notes*, 4(265), 2011.

[PH11] Holger Pillmann, Klas Hatje, Florian Odronitz, Björn Hammesfahr, and Martin Kollmar. Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC Bioinformatics*, 12(270), 2011.

Trans-species transcriptome analysis to identify novel lifespan-affecting genes

The JenAge Consortium*

* Leibniz Institute for Age Research – Fritz Lipmann Institute, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Friedrich Schiller University, University Hospital; Jena, Germany

Ageing is a complex biological phenomenon involving changes of gene expression, metabolic and signaling pathways, which ultimately result in a progressive loss of function and death. It is now being increasingly recognized that a reductionist approach ascribing ageing phenomena to single causes is inadequate to explain all ageing-related functional changes. Therefore, the potential value of applying systems biological approaches to foster our understanding of these processes is widely recognized. The Jena Centre for Systems Biology of Ageing – JenAge – aims at a multi-species approach to study the impact of mild stress on healthy ageing (http://www.jenage.de).

Towards this goal we set out to systematically characterize the changes of transcriptomes for five species ranging from *Caenorhabditis elegans* over two fish models and mice to man at different ages. Data have been acquired by RNA-seq using the Solexa/Illumina platform. First results characterizing the process of normal ageing within and between species by fuzzy c-means clustering, gene set enrichment analysis, machine learning of non-linear models by induction of decision trees as well as correlation network inference will be presented.

As a proof of principle, we identified genes that were found to be uniformly up- or down-regulated with age in *C. elegans*, zebrafish and mouse. Next, worms were cultivated on bacteria containing candidate-specific RNAi constructs and lifespan was determined. Remarkably, the knock-down of the majority of the candidate genes led to alterations of lifespan in *C. elegans*, thus underscoring the suitability of our approach.

Comparison of high-throughput technologies of Aspergillus fumigatus using RNA-Seq

Sebastian Müller, Marco Groth, Konrad Grützmann, Reinhard Guthke,Olaf Kniemeyer, Axel A. Brakhage, Vito Valiante

Research Group Systems Biology / Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knoell Institute, Jena
sebastian.mueller@hki-jena.de

The filamentous fungus Aspergillus fumigatus has become the most important airborne fungal pathogen causing life-threatening infections in immunosuppressed patients. Recent developments in high-throughput technologies to measure the transcriptome or proteom such as microarrays, RNA-Seq and 2D DIGE are getting more and more popular to investigate organisms systematically. In particular for investigating the transcriptome, the question which one to take arises naturally, since a lot of alternatives are at hand. Here, we conduct for the first time a comprehensive comparison of 5 microarray studies based 4 different platform, 2 RNA-Seq studies from different laboratories as well as a 2D DIGE proteom study. Further we highlight the potential for paired-end RNA-Seq data to investigate the Aspergillus transcriptome, applying it to the Wild-type and the delta−mpkA mutant. We were able to identify a substantial number of novel transcripts, detecting hundreds of SNP's, new exons, untranslated regions, thousands of new splice junctions and detected widespread alternative splicing events. Many genes and gene clusters such as the Gliotoxin cluster, the Pseurotin A cluster or a cluster with a still unknown product were found to be differentially regulated.

References

[SM12] Müller S, Baldin C, Groth M, Guthke R, Kniemeyer O, Brakhage AA, Valiante V: **Comparison of transcriptome technologies in the pathogenic fungus Aspergillus fumigatus reveals novel insights in genome structure and MpkA-dependent gene expression**. *BMC Genomics*, in revision

Visualization of Protein Ligand Graphs

Tim Schäfer, Ina Koch
Institute of Computer Science, Department of Molecular Bioinformatics, Goethe-University Frankfurt, Robert-Mayer-Straße 11–15, 60325 Frankfurt am Main, Germany
tim.schaefer, ina.koch@bioinformatik.uni-frankfurt.de

Ligand information is of great interest to understand protein function. Protein structure topology can be modeled as a graph with secondary structure elements as vertices and spatial contacts between them as edges. Meaningful representations of such graphs in 2D are required for the visual inspection, comparison and analysis of protein folds, but their automatic visualization is still challenging. We present an approach which solves this task, supports different graph types and can optionally include ligand contacts.

Our method extends the field of protein structure description and visualization by including ligand information. It generates a mathematically unique representation and high-quality 2D plots of the secondary structure of a protein based on a protein-ligand graph. This graph is computed from 3D atom coordinates in PDB files and the corresponding SSE assignments of the DSSP algorithm [KS83]. The related software supports different notations and allows a rapid visualization of protein structures. It can also export graphs in various standard file formats so they can be used with other software. Our approach visualizes ligands in relationship to protein structure topology and thus represents a useful tool for exploring protein structures.

The software is released under an open source license and available at `http://www.bioinformatik.uni-frankfurt.de/` in the *Software* section under *Visualization of Protein Ligand Graphs*.

References

[KS83] W. Kabsch and C. Sander. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, 22:2577–2637, 1983.

Comparative Genomics in the Social Amoebae

Andrew J. Heidel[1], Hajara M. Lawal[2], Marius Felder[1], Christina Schilde[2], Nicholas R. Helps[2], Budi Tunggal[3], Francisco Rivero[4], Uwe John[5], Michael Schleicher[6], Ludwig Eichinger[3], Matthias Platzer[1], Angelika A. Noegel[3], Pauline Schaap[2], Gernot Glöckner[1,3,7]

[1] Leibniz Institute for Age Research – Fritz Lipmann Institute, [2] College of Life Sciences, University of Dundee, [3]University of Cologne, [4] Hull York Medical School and Department of Biological Sciences University of Hull, [5] Alfred Wegener Institute, [6] Institute for Anatomy and Cell Biology, and Center for Integrated Protein Science (CIPSM),Ludwig-Maximilians-University Munich, [7] Leibniz-Institute of Freshwater Ecology and Inland Fisheries

aheidel@fli-leibniz.de

Dictyostelium discoideum (DD), an extensively studied model organism for cell- developmental- and evolutionary biology, belongs to the most derived group 4 of social amoebas, a clade of altruistic multicellular organisms. To understand the common genetic basis of social evolution and define species-specific adaptations, we completely sequenced with the help of next-generation sequencing, the genomes of two species, *D. fasciculatum* (DF) and *P. palllidum* (PP), that represent the basal groups 1 and 2 of social amoebas. In combination with the already sequenced *D. discoideum* these species cover the breadth of the social amoebae lineage. The number of protein coding genes is similar between species, but only half of them comprise an identifiable set of orthologous genes. In general, genes involved in primary metabolism, cytoskeletal functions and signal transduction are conserved, while genes involved in secondary metabolism, export and signal perception underwent large differential gene family expansions. This most likely signifies involvement of the conserved set in core cell- and developmental mechanisms, and of the diverged set in niche- and species-specific adaptations for defence and food, mate- and kin selection. Several indicators such as protein divergence, high ratio of non-synonymous to synonymous mutations and extensive loss of synteny show that *DF*, *PP* and *DD* split from their last common ancestor at least 0.6 billion years ago.

References

[Knu84] Donald E. Knuth. The TEXbook. In *Computers and Typesetting*, volume A. Addison-Wesley, Massachusetts, second edition edition, 1984.

Eukaryotic Gene Prediction Maximizing Posterior Accuracy

Lizzy Gerischer, Mario Stanke
Institute of Mathematics and Computer Science
University of Greifswald
lizzy.gerischer@uni-greifswald.de

Ab initio gene prediction programs are commonly based on a probabilistic model of all possible gene structures and predict a gene structure with highest probability in this model. It is often a hidden Markov model (HMM) or a conditional random field and the algorithm for gene structure inference is then some variant of the Viterbi algorithm. Such an approach tries to minimize the probability of having *any difference* between the predicted and the correct gene structure. However, it does not consider any similarity between these two gene structures *if they differ*, e.g. that almost - but not quite - correct gene structures are more useful than completely missing genes. We therefore examined an approach that takes the degree of similarity of gene structures into account as done with multiple sequence alignments in [DBBM05].

Hence, we are interested in the gene structure that maximizes the similarity to the correct gene structure (accuracy). Since no information about the correct gene structure is given, we maximize the expected accuracy. The problem of finding the most accurate gene structure was solved by transforming it to a "shortest path" problem in directed graphs.

References

[DBBM05] Chuong B. Do, Michael Brudno, Serafim Batzoglou, and Mahathi S.P. Mahabhashyam. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15:330–340, 2005.

The transcript catalogue of the short-lived fish *Nothobranchius furzeri* provides insights into age-dependent changes of mRNA levels

Andreas Petzold[1], Kathrin Reichwald[1], Marco Groth[1], Stefan Taudien[1], Nils Hartmann[2], Steffen Priebe[3], Dmitry Shagin[4], Christoph Englert[2] and Matthias Platzer[1]

1 Genome Analysis, 2 Molecular Genetics, Leibniz Institute for Age Research – Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany

3 Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute, Beutenbergstr. 11a, 07745 Jena, Germany

4 Evrogen Ru SJC, 5 Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Milukho-Maklaya 16/10, Moscow, 117997, Russia

andpet@fli-leibniz.de

The African annual fish *Nothobranchius furzeri* has over recent years been established as a model species for ageing-related studies, which has mainly been based on its exceptionally short lifespan and the presence of typical characteristics of vertebrate ageing. [Gen05] To substantiate its role as an alternative vertebrate ageing model, a transcript catalogue is needed, which can serve e.g. as basis for identifying ageing-related genes.

To build the *N. furzeri* transcript catalogue, thirteen cDNA libraries were sequenced using Sanger, 454/Roche and Solexa/Illumina technologies yielding about 39 Gb. In total, 19,875 protein-coding genes were identified and annotated. Of these, 71% are represented by at least one transcript contig with a complete coding sequence. Further, transcript levels of young and old fish of the strains GRZ and MZM-0403, which differ in lifespan by 100%, were studied by RNA-seq. Eighty-six differentially expressed genes were detected in skin and brain; these have a role in cell cycle and proliferation, inflammation and tissue maintenance. An RNA-seq experiment for zebrafish skin confirmed the ageing-related relevance of the findings in *N. furzeri*. Notably, analyses of transcript levels between *N. furzeri* strains as well as zebrafish differed largely, suggesting that ageing is accelerated in the short-lived *N. furzeri* strain GRZ compared to the longer-lived strain MZM-0403.

We provide a comprehensive, annotated *N. furzeri* transcript catalogue and a first transcriptome-wide insight into *N. furzeri* ageing. This data will serve as a basis for future functional studies of ageing-related genes.

References

[Gen05] Genade T, Benedetti M, Terzibasi E, Roncaglia P, Valenzano DR, Cattaneo A, Cellerino A: Annual fishes of the genus Nothobranchius as a model system for aging research. *Aging Cell* 2005, 4:223–233.

Quantitative Model of Cell Cycle Arrest and Cellular Senescence in Human Fibroblasts

S. Schäuble[1,2], K. Klement[3], S. Marthandan[3], S. Münch[3], I. Heiland[2], S. Schuster[2], P. Hemmerich[3], S. Diekmann[3]
[1]*Research Group Theoretical Systems Biology, Friedrich-Schiller-University, Jena, Germany*
[2]*Dept. of Bioinformatics, Friedrich-Schiller-University, Jena, Germany*
[3]*Fritz-Lipmann-Institute, Jena, Germany*
sascha.schaeuble@uni-jena.de

Primary human fibroblasts undergo a limited number of cell divisions before entering a non-replicative senescent state [HM61]. While at early population doublings fibroblasts are proliferation-competent (P), an increasing number of cells become reversibly cell cycle arrested (C) and finally irreversibly senescent (S) during further cell passaging. We have developed a new quantitative model of this stepwise transition, including a stress response function F that aggregates and processes various forms of stress [SKM$^+$12]. Applying senescence marker quantification to our model, allowed us to discriminate between the cellular states P, C, and S as well as to identify the transition rates between them for different human fibroblast cell types. Unexpectedly, our model-derived quantification revealed significant differences in the stress response of different fibroblast cell lines. We found that SA-β-Gal is a good quantitative marker for cellular senescence in WI-38 and BJ cells, but much less so in MRC-5 cells. The differentiation between three cellular states and the explicit separation of stress induction from the cellular stress response allows us for the first time to quantitatively assess the response of primary human fibroblasts towards endogenous and exogenous stress during cellular ageing.

References

[HM61] L. Hayflick and P. S. Moorhead. The serial cultivation of human diploid cell strains. *Exp Cell Res*, 25:585–621, Dec 1961.

[SKM$^+$12] Sascha Schäuble, Karolin Klement, Shiva Marthandan, Sandra Münch, Ines Heiland, Stefan Schuster, Peter Hemmerich, and Stephan Diekmann. Quantitative model of cell cycle arrest and cellular senescence in primary human fibroblasts. *PLoS One*, 7(8):e42150, 2012.

An Effective Framework for Reconstructing Gene Regulatory Networks From Genetical Genomics Data

R. J. Flassig, S. Heise, K. Sundmacher and S. Klamt
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany
heise@mpi-magdeburg.mpg.de

Systems Genetics approaches, in particular those relying on genetical genomics data, put forward a new paradigm of large-scale genome and network analysis. These methods use naturally occurring multifactorial perturbations (e.g. polymorphisms) in properly controlled and screened genetic crosses to elucidate causal relationships in biological networks. However, although genetical genomics data contain rich information, a clear dissection of causes and effects as required for reconstructing gene regulatory networks is not easily possible. We present a framework for reconstructing gene regulatory networks from genetical genomics data where genotype and phenotype correlation measures are used to derive an initial graph which is subsequently reduced by pruning strategies to minimize false positive predictions [1]. Applied to realistic simulated genetic data from a recent DREAM challenge we demonstrate that our approach is simple yet effective and outperforms more complex methods (including the best performer) with respect to (i) reconstruction quality (especially for small sample sizes) and (ii) applicability to large data sets due to relatively low computational costs.

References

[1] Klamt, S. et al. TRANSWESD: inferring cellular networks with transitive reduction. *Bioinformatics*, 26: 2160-2168, 2010

Analysis of evolutionary constraints on transcription factor binding sites in Arabidopsis thaliana

Paula Korkuć and Dirk Walther
Max Planck Institute of Molecular Plant Physiology, Potsdam University
{korkuc|walther}@mpimp-golm.mpg.de

One of the central interests in modern biology is the identification of functional elements encoded in a genome via comparative genomics. We exploited the genomic sequencing information of a high number of different accessions of model plant Arabidopsis thaliana as available from the 1001 genome project to characterize known and identify novel regulatory elements in gene promoter regions of Arabidopsis. Assuming that promoter regions and regulatory elements such as transcription factor binding sites (TFBSs) tend to be more conserved than non-functional intergenic regions, we wanted to estimate the bounds of promoter regions by determining the density of single nucleotide polymorphisms (SNPs) along the intergenic regions, verify known TFBSs by analyzing their localization versus their level of conservation with respect to the transcription start site, and find new potential motifs out of all possible DNA hexameres. Based on the obtained SNP density profile, the average length of promoter regions could be established at 500 nt. We confirmed that known TFBS-motifs are indeed more conserved than the promoter background ($p = 2.2e^{-16}$). For eight known motifs their positional preferences could be clearly substantiated based on their position-specific SNP density. Lastly, nine new candidate motifs were identified whose relative positional occurrence correlates highly with their level of position-specific conservation. For those candidate motifs, experimental verification will be required. Our study demonstrates that the now available resolution of SNP data offers novel ways for the identification of functional elements and the characterization of gene promoter sequences in general.

Predicting protein interfaces by modeling spatial structures

Torsten Wierschin, Mario Stanke
Institut für Mathematik und Informatik, Universität Greifswald
{torsten.wierschin,mario.stanke}@uni-greifswald.de

Predicting interaction sites on the surface of proteins remains an active research area. Applications like antibiotic drug design and prediction of protein docking processes would gain substantial advantages from precise information about protein-protein interfaces. Previous work includes the classification of the residues with machine learning approaches. Thereby the problem is formulated as to assign a label to each residue being in the interface or not. [LLWL07] considered some interdependencies between labels of different residues by interpreting the problem as one of sequentially labeling the residues of the protein sequence and using a linear-chain conditional random field (CRF). Our graphical model reflects the assumption that the label of one residue is conditionally independent of the labels of residues further than a distance threshold, *given* the labels of the other residues closer than the threshold ($3\mathring{A} - 12\mathring{A}$). This assumption is weaker than the independence assumption that underlies a sequential model, which does not consider dependencies between residues that are spatially close but not in sequence. Although we currently consider only the feature *relative surface accessibility*, we obtain better classification results than Li et al. who use several more features. CRF parameters were estimated using a modification of the online large margin algorithm, [BCHP07].

References

[BCHP07] Axel Bernal, Koby Crammer, Artemis Hatzigeorgiou, and Fernando Pereira. Global Discriminative Learning for Higher-Accuracy Computational Gene Prediction. *PLoS Comput Biol*, 3(3):e54, 2007.

[LLWL07] Ming-Hui Li, Lei Lin, Xiao-Long Wang, and Tao Liu. Protein-protein interaction site prediction based on conditional random fields. *Bioinformatics*, 23(5):597–604, 2007.

Theoretical design and experimental verification of amino acid overproducing strains of *Escherichia coli* using CASOP GS

Silvio Waschina[1*], Christoph Kaleta[1], Christian Kost[2]
[1] *Research Group Theoretical Systems Biology, Friedrich-Schiller-University Jena*
[2] *Department of Bioorganic Chemistry, Max Planck Institute for Chemical Ecology*
[*]silvio.waschina@uni-jena.de

Genome-scale computational tools are able to identify genetic targets for strain improvement of amino acid overproducing microorganisms. CASOP GS (a Computational Approach for Strain Optimization aiming at high Productivity - Genome-Scale) is an algorithm for the prediction of reaction knock-out and overexpression strategies for metabolic engineering based on genome-scale metabolic networks [HK10]. It takes advantage of the theory of elementary flux modes (EFMs). In contrast to other approaches for in silico target identification, the predictions made by CASOP GS do not only incorporate product yield, but also a combined measure of conversion capacity and yield.

Escherichia coli BW25113 was engineered for the production of L-tryptophan. To this end, reaction knock-out targets in the metabolic network were identified using CASOP GS. Associated genes of these targets were deleted and the production of all amino acids (except L-cysteine) was measured by LC/MS/MS. Five single- and six double gene deletion mutants were generated. Two double gene deletion mutants ($\Delta purU\ \Delta ppc$ and $\Delta pykA\ \Delta ppc$) showed a 4-fold increase in their tryptophan production relative to the wild type. Interestingly, the double gene deletion mutants also overproduced most other amino acids, whereas single gene deletion mutants did not show amino acids overproduction.

This study showed the potential of systems metabolic engineering, where an iterative cycle of theoretical predictions and experimental implementation provides a rational basis for the targeted construction of strains overproducing certain metabolites of interest.

Reference

[HK10] Oliver Hädicke and Steffen Klamt. CASOP: a computational approach for strain optimization aiming at high productivity. *J Biotechnol*, 147(2):88–101, May 2010.

Supervised Penalized Canonical Correlation Analysis

Andrea Thum, Lore Westphal, Tilo Lübken, Sabine Rosahl, Steffen Neumann, Stefan Posch
Institute of Computer Science, Martin-Luther-University Halle-Wittenberg
andrea.thum@informatik.uni-halle.de

The canonical correlation analysis (CCA) is commonly used to analyze relationships between features of data sets with paired data, e.g., measurements of gene expression and metabolite intensities of the same experiments. The CCA tries to find linear combinations of the features in each data set, that correlate maximally: the canonical variables. These combinations correspond to processes within the organism. For data sets with more features than samples, standard CCA is not applicable. Instead, a penalized CCA ([WZ09]) can be used.
However, it can be difficult to interpret the underlying processes: the relationship is not easy to infer and often the relationships observed are not related to the experimental design but to some unknown parameters.

Here we present an extension of the penalized CCA, the *supervised penalized* CCA. The experimental design is used as a third data set. The correlations of the canonical variables of the biological data sets and this *design data set* are maximized to find interpretable and meaningful relationships. To this end a generalized CCA for several data sets ([VSP06]) is modified to include the penalty parameters.
The supervised pCCA was successfully tested on a data set of *Arabidopsis thaliana* with gene expression and metabolite intensity measurements. It resulted in seven significant canonical variables and their interpretation.

References

[VSP06] J. Via, I. Santamaria, and J. Perez. A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Networks*, 20:139–152, 2006.

[WZ09] S. Waaijenborg and A. H. Zwinderman. Correlating multiple SNPs and multiple disease phenotypes: penalized non-linear canonical correlation analysis. *Bioinformatics*, 25 (21):2764–2771, 2009.

An Efficient Data Structure for Pangenomes

Corinna Ernst and Sven Rahmann
Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, Germany
corinna.ernst@uni-due.de, sven.rahmann@uni-due.de

Due to the increasing amount of sequence data current biological research focusses more and more on the exploration of the global gene repertoire of related species, referred to as the pangenome, instead of single genomic sequences. However, an efficient and variable data structure representing the pangenome of a given set of genomes or genomic strains is not available yet.
Following the concept of pangenomes, genomic regions are categorized according to their prescence in the entire set of available genomes [TMC+05]. Once a subsequence is known to appear slightly modified in another genomic region, its sequence can be efficiently described by a progression of edit operations applied to this region [BWB09]. We present here a data structure for pangenomes based on pooling shared genomic features into a single object, consisiting of a common reference sequence and providing for each entry the edit operations required for sequence retrieval. Beside straightforward access to pangenome-related information, dependent on similarity between the input sequences, our data structure is expected to reduce storage capacity compared to raw sequences by avoiding to keep redundant information. Furthermore, it is variable in the sense that pangenome-related information or even entire genomes may be removed or added at any time. Locations of shared genomic regions can be included easily by parsing *GenBank* and *BLAST* or *nucmer* output.

References

[BWB09] Marty C. Brandon, Douglas C. Wallace, and Pierre Baldi. Data structures and compression algorithms for genomic sequence data. *Bioinformatics*, 25(14):1731–1738, 2009.

[TMC+05] Herv Tettelin, Vega Masignani, Michael J Cieslewicz, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955, 2005.

ISOQuant - an integrated bioinformatics pipeline for evaluation and reporting of data independent (LC-MSE) label-free quantitative proteomics data

Jörg Kuharev, Hansjörg Schild and Stefan Tenzer
UMC of the Johannes-Gutenberg-University Mainz, Mainz, Germany,
kuharev@uni-mainz.de

One of the main bottlenecks in the evaluation of label-free quantitative proteomics experiments is the lack of integrated software solutions and the often cumbersome export of processed data for in-depth evaluation and comparative analysis. Recent developments facilitate the reproducible detection and quantification of up to two thousand proteins within a single 1D LC-HDMSE experiment. However, data-independent, alternate scanning LC-MS peptide fragmentation data can currently only be processed by vendor software, which is limited to analysis on a run-by-run basis. We present the bioinformatics pipeline ISOQuant - an integrated solution for automated post-processing in-depth evaluation of label-free LC-MS data, allowing easy data access and export to common third party formats. Vendor software (PLGS[1]) is used for raw data processing and for peptide and protein identification. ISOQuant automatically extracts LC-MS experiment data from vendor software and imports relevant information into a relational database (MySQL[2]) subsequently applying a set of in-house developed and adapted third party analysis methods. Relations between multiple LC-MS runs are built and advanced statistics calculated. Non-linear retention time distortions between LC-MS runs are corrected[3]. Corresponding signals are clustered and subjected to multidimensional intensity normalization. Clusters are annotated by consensus peptides from associated LC-MS runs. Homologue proteins are filtered. Shared peptide intensities are redistributed. Absolute in-sample amounts are calculated[4]. Finally, results of the performed analysis are exported as a set of uniform reports. ISOQuant provides easy access to routine application of label-free quantification by significantly reducing evaluation time and by offering standardized data evaluation procedures. ISOQuant and related resources are freely available at http://www.isoquant.net/.

References

1. Waters: ProteinLynx Global SERVER (PLGS). at http://www.waters.com/waters/nav.htm?cid=513821
2. MySQL: The world's most popular open source database. at http://www.mysql.com/
3. Podwojski, K. *et al*. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* **25**, 758 –764 (2009).
4. Silva, J. C. *et al*. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell Proteomics* **5**, 144–156 (2006).

On the evolutionary significance of the size and planarity of the proline ring

Jörn Behre[1,4], Roland Voigt[3], Ingo Althöfer[2] and Stefan Schuster[1]

[1]Dept. of Bioinformatics and [2]Dept. of Mathematics,
Friedrich Schiller University,
Ernst-Abbe-Platz 2, 07743 Jena, Germany
[3]Institute of Mathematics, University of Leipzig,
Johannisgasse 26, 04103 Leipzig, Germany
[4]Institute of Food Research
Colney Lane, Norwich NR4 7UA, U.K.
Phone +49-3641-949580, Fax +49-3641-946452
stefan.schu@uni-jena.de

In biological evolution, from an enormous multitude of possibilities, a much smaller set of solutions is selected. For example, from all aliphatic amino acids possible, only a few are used in proteins [Grü11]. Proline is a proteinogenic amino acid in which the side chain forms a ring, the pyrrolidine ring. This is a five-membered ring made up of four carbons and one nitrogen. Here, we study the evolutionary significance of this ring size, in comparison to other conceivable ring sizes. It is shown that the pyrrolidine ring has the advantage of being nearly planar and strain-free, based on a general mathematical assertion saying that the angular sum of a polygon is maximum if it is planar and convex [Beh12]. The optimality of the ring size of proline can be derived from a triangle inequality for angles. Quasi-planarity is physiologically significant because it allows an easier and evolutionarily old type of fit into binding grooves of proteins with which proline-rich proteins interact. Finally, we present a comparison with other planar, nearly planar and non-planar biomolecules such as neurotransmitters, hormones and toxins, involving, for example, aromatic rings, cyclopentanone and 1,3-dioxole [Beh12].

References

[Beh12] J. Behre, R. Voigt, I. Althöfer, S. Schuster: On the evolutionary significance of the size and planarity of the proline ring. *Naturwissenschaften*, in press.

[Grü11] K. Grützmann, S. Böcker, S. Schuster: Combinatorics of aliphatic amino acids. *Naturwissenschaften* 98: 79-86, 2011.

rBiopaxParser: A new package to parse, modify and merge BioPAX-Ontologies within R

Frank Kramer*, Michaela Bayerlová, Annalen Bleckmann, Tim Beissbarth
Department of Medical Statistics, University Medical Center Göttingen, Göttingen
frank.kramer@med.uni-goettingen.de

Methods for network reconstruction are often designed with the possibility to integrate prior knowledge about the topology of biological signaling networks. In the past years ontologies have been the tool of choice to represent and allow the sharing of knowledge of this biological reality. BioPAX is a commonly used ontology for the encoding of regulatory pathways [DE10]. The R Project for Statistical Computing is the standard environment for statistical analyses of high-dimensional data and network reconstruction methods. Although there are packages available that provide the pathway data of databases like KEGG, the Pathway Interaction Database (Nature/NCI) or Reactome as graphs, there was no software available to parse, merge and manipulate BioPAX ontologies inside of R. We present a new open-source package called rBiopaxParser that parses BioPAX-Ontologies and represents them in R. The user is able to parse arbitrary BioPAX OWL files, for example, the exports of popular online pathway databases like PID, Reactome or KEGG. Instances of BioPAX-Classes can be programatically added or removed. Multiple pathways can be merged or transformed into an adjacency matrix suitable as input for network reconstruction algorithms, i.e. reducing a pathway to a graph with edges representing only activations or inhibitions. Introductory vignettes as well as extensive documentation are available online and as R Help. The software is freely available at https://github.com/frankkramer/rBiopaxParser and will be submitted to Bioconductor.

References

[DE10] Emek Demir, Gary D. Bader, et al. The BioPAX community standard for pathway data sharing. *Nat Biotechnol.*, 2010 Sep;28(9):935-42.

Visualizing Peptide Spectrum Matches in Genome Browsers

Mathias Kuhring and Bernhard Y. Renard
Research Group Bioinformatics (NG4),
Robert Koch-Institute, Berlin, Germany
RenardB@rki.de

Proteogenomic approaches have gained increasing popularity, however it is still difficult to integrate mass spectrometry identifications with genomic data due to differing data formats. To address this difficulty, we developed the "integrating Peptide spectrum matches into Genome browser visualizations" (iPiG) tool. Thereby, the concurrent analysis of proteomic and genomic data is significantly simplified and proteomic results can directly be compared to genomic data. The main idea of iPiG is the mapping of peptide spectrum matches (PSMs) [NVA07] to their corresponding gene of origination using a comprehensive set of known genes. The mapping requires a matching of protein identifiers to gene identifiers as well as a string matching of the PSM peptide sequence to the translation of the selected gene. This results in the exact genomic position of the origin of the PSM provided by the location annotation of the gene. Unmapped PSMs are treated again with an overall string matching search to the gene set. Finally, the genomic positions of all mapped PSMs are exported as genome annotation tracks in formats such as bed or gff3. Those formats can easily be loaded into most common genome browser such as the UCSC Genome Browser [KSF$^+$02]. Once imported into a genome browser, the PSMs are automatically visualized at the right positions by the browser. This allows a genome position-oriented and gene expression-like overview of peptide spectrum matches. iPiG is implemented in Java with a graphical user interface. It is freely available from https://sourceforge.net/projects/ipig/.

References

[KSF$^+$02] W. James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, Haussler, and David. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002.

[NVA07] Alexey I. Nesvizhskii, Olga Vitek, and Ruedi Aebersold. Analysis

and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods*, 4(10):787–797, January 2007.

Detecting and investigating substrate cycles in a genome-scale human metabolic network

Juliane Gebauer[1,2], Stefan Schuster[1], Lus F. de Figueiredo[1,3] and Christoph Kaleta[1,2]
[1] *Dept. of Bioinformatics, Friedrich Schiller University of Jena*
[2] *Research Group Theoretical Systems Biology, Friedrich Schiller University of Jena*
[3] *Current address: Cheminformatics and Metabolism, European Bioinformatics Institute (EBI)*
Juliane.Gebauer@uni-jena.de, Christoph.Kaleta@uni-jena.de

Substrate cycles, also known as futile cycles, are cyclic metabolic routes that dissipate energy by hydrolysing cofactors such as ATP. A popular example is the conversion of fructose-6-phosphate to fructose-1,6-bisphosphate and back. We analyse a large number of substrate cycles in human metabolism, which are consuming ATP and discuss their statistics [GSdFK12]. For this purpose we use two recently published methods, EFMEvolver [KdFBS09] and the K-shortest EFM method [dFPR$^+$09], to calculate samples of 100,000 and 15,000 substrate cycles, respectively. We find a surprisingly high number of substrate cycles in human metabolism, with up to one hundred reactions per cycle, utilizing reactions from through up to six different compartments. An analysis of tissue specific models of liver and brain metabolism shows that there is selective pressure acting against the uncontrolled dissipation of energy by avoiding the coexpression of enzymes belonging to the same substrate cycle.

References

[dFPR$^+$09] Luis F. de Figueiredo, Adam Podhorski, Angel Rubio, Christoph Kaleta, John E. Beasley, Stefan Schuster, and Francisco J. Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25(23):3158–3165, Dec 2009.

[GSdFK12] Juliane Gebauer, Stefan Schuster, Lus F. de Figueiredo, and Christoph Kaleta. Detecting and investigating substrate cycles in a genome-scale human metabolic network. *FEBS J*, Jul 2012.

[KdFBS09] Christoph Kaleta, Luis F. de Figueiredo, Joern Behre, and Stefan Schuster. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. 157:180–190, 2009.

Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes

Janine Roy, Christof Winter, Zerrin Isik, Michael Schroeder
BIOTEC, Technische Universität Dresden
Tatzberg 47-49, 01307 Dresden
janine.roy@biotec.tu-dresden.de

In the last decade, there has been much work on predicting disease progression and other outcome variables from gene expression to personalize treatment options. Despite first diagnostic kits on the market, found marker genes often show limited prediction accuracy, limited reproducibility, and unclear biological relevance. In order to solve these problems, we developed a novel outcome prediction algorithm -NetRank- to identify marker genes prognostic for outcome using both expression data and network information. Our approach adapts the random surfer model of Google's PageRank algorithm to rank genes according to their prognostic relevance. We applied the algorithm to gene expression profiles obtained from 30 pancreas cancer patients, and identified seven candidate marker genes. Compared to genes found with state of the art methods, NetRank improved the prediction accuracy by 7%. When experimentally validating the prognostic value of our seven candidate markers on an independent set of 412 pancreatic cancer samples, we achieved an accuracy superior to established clinical prognostic factors [WKK+12]. Besides, we systematically evaluated the prognostic power of networks and NetRank for signature identification on 25 published cancer datasets. We could show that NetRank algorithm performs better than classic feature selection methods. In addition, reproducibility of signatures created by NetRank significantly increases between different datasets of the same cancer type. In future network-based gene expression analysis will lead to a more detailed understanding of cancer-related processes.

References

[WKK+12] Winter, C. et al. (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol*, 8(5):e1002511, May 2012

Give it AGO! - Insights into microRNA Argonaute sorting in plants

Christoph Thieme and Dirk Walther
Max Planck Institute for Molecular Plant Physiology, Potsdam
thieme@mpimp-golm.mpg.de

MicroRNAs (miRNA) are a class of non-coding RNAs which mediate a variety of important functions by post-transcriptional regulation of specific target transcripts. After maturation from double-stranded precursors, short 20 to 24nt miRNA sequences are recognized by the RNA-induced silencing complex (RISC). High sequence complementarity between the miRNA and the target mRNA allows for highly selective binding of the target by the RISC.

The main component of the RISC complex are proteins from the Argonaute (AGO) family. The *Arabidopsis thaliana* genome encodes ten different AGOs. Observations show that different AGOs preferentially bind specific miRNAs and previous studies indicate the importance of the $5'$ nucleotide of the miRNA for this AGO-sorting process. However, this fairly simple rule has several exceptions and appears to only allow four different AGOs to be selectively recognized.

We used a bioinformatics approach to investigate additional features that may possibly be relevant for AGO sorting. Based on published RNA high-throughput sequencing data isolated by AGO crosslinking immunoprecipitation (HITS-CLIP data) we extracted 148 miRNAs specifically associated with either AGO1, AGO2 or AGO5. On these miRNA-AGO pairs we performed random forest classification and mutual information analysis. Furthermore, we applied clustering methods to microarray data of AGO expression to assess the impact of AGO presence and absence on miRNA binding.

We found that AGO-sorting relies not only on the $5'$ nucleotide of the miRNA but to a certain degree also on several signals of the mature sequence as well as on the match-mismatch pattern of the precursor structure. Additionally, spatial and temporal presence of the different AGO proteins appear to be important factors for the fate of miRNAs.

Modeling the Seasonal Adaptation of Circadian Clocks by Changes in the Network Structure of the Suprachiasmatic Nucleus

Christian Bodenstein[1], Marko Gosak[2], Stefan Schuster[1], Marko Marhl[2] and Matjaž Perc[2]

[1]*Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany*

[2]*Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia*

christian.bodenstein@uni-jena.de, marko.gosak@uni-mb.si

The dynamics of circadian rhythms needs to be adapted to day length changes between summer and winter. It has been observed experimentally, however, that the dynamics of individual neurons of the suprachiasmatic nucleus (SCN) does not change as the seasons progress. Rather, the seasonal adaptation of the circadian clock is assumed to be the consequence of changes in the intercellular dynamics, which leads to a phase distribution of electrical activity of SCN neurons that is narrower in winter and broader during summer. Yet to understand this complex intercellular dynamics, a more thorough understanding of the impact of the network structure formed by the SCN neurons is needed. To that effect, we propose a mathematical model for the dynamics of the SCN neuronal architecture in which the structure of the network plays a pivotal role. We show that the fraction of long-range cell-to-cell connections and the seasonal changes in the daily rhythms are tightly related. In particular, simulations of the proposed mathematical model indicate that the fraction of long-range connections between the cells adjusts the phase distribution and consequently the length of the behavioural activity as follows: dense long-range connections during winter lead to a narrow activity phase, while rare long-range connections during summer lead to a broad activity phase. Our model is also able to account for the experimental observations indicating a larger light-induced phase-shift of the circadian clock during winter, which we show to be a consequence of higher synchronization between neurons. We thus provide evidence that the variations in the seasonal dynamics of circadian clocks can be understood and regulated by the plasticity of the SCN network structure.

Genome sequence analysis of three marine fungal isolates using different next generation DNA sequencing methods

Abhishek Kumar and Frank Kempken

Abteilung für Botanik mit Schwerpunkt Genetik und Molekularbiologie,
Botanisches Institut, Christian-Albrechts-Universität zu Kiel, Germany.
akumar@bot.uni-kiel.de | fkempken@bot.uni-kiel.de

Enormous biodiversity of marine fungal isolates is mirrored by the molecular diversity of their secondary metabolites [Kon06]. Over 100 terrestrial fungal genomes using Sanger method have been sequenced, and recently the *Sordaria macrospora* genomic sequences became available using next-generation sequencing [Now10]. However, so far, marine isolates of fungi are lacking genetic information to unravel their genetic properties. Here, we report the first study of genomic and RNA sequencing for three marine fungal isolates namely, *Scopulariopsis brevicaulis, Pestalotiopsis* sp. and *Calcarisporium* sp. *S. brevicaulis* is known to produce the cyclic peptides Scopulariде A and B [Zni2008]. *Pestalotiopsis* sp. and *Calcarisporium* sp. are endophytic fungi and they produce wide array of secondary metabolites. We have established the genomic sequences from these marine isolates of using three different next-generation sequencing methods (Roche 454, Illumina and ion-torrent) and predicted genes are presently in process of validation using illumina based RNA-seq. We present our current data for *S. brevicaulis*, *Pestalotiopsis* sp. and *Calcarisporium* sp. assembled genome of about 32 Mb (935 contigs/N50 – 88 kb), 46 Mb (4186 contigs/N50 – 71 kb), and 36 Mb (2464 contigs/N50 – 90.8 kb), respectively. We will also present our approach for genome assembly and annotation analyses. Furthermore, we will provide results of secondary metabolite encoding gene profiles. In the EU-funded project marine fungi (www.marinefungi.eu), isolations and characterisations of new anti-cancer compounds from these fungi are underway.

References

[Kon06] König et al. Natural Products from Marine Organisms and their Associated Microbes. Chembiochem 7(2):229-38, 2006.

[Now10] Nowrousian et al. De novo Assembly of a 40 Mb Eukaryotic Genome from Short Sequence Reads: Sordaria macrospora, a Model Organism for Fungal Morphogenesis. PLoS Genet 6(4): e1000891, 2010

[Zni2008] Zhiguo et al. Scopularides A and B, Cyclodepsipeptides from a Marine Sponge-Derived Fungus, Scopulariopsis brevicaulis. Journal of Natural Products 71 (6), 1052-1054, 2008.

Sequence, structure, function and evolution of BEM46 proteins

Abhishek Kumar, Krisztina Kollath-Leiß and Frank Kempken

Abteilung für Botanik mit Schwerpunkt Genetik und Molekularbiologie,
Botanisches Institut, Christian-Albrechts-Universität zu Kiel, Germany.
akumar@bot.uni-kiel.de | fkempken@bot.uni-kiel.de

The bud emergence 46-like (BEM46) protein from *Neurospora crassa* belongs to the alpha/beta-hydrolase superfamily. Recently, we have reported that the BEM46 protein is localized in the perinuclear ER and also forms spots close by the plasma membrane [Mer09]. The protein appears to be required for cell type-specific polarity in *Neurospora crassa*. Furthermore, initial studies suggested that the BEM46 amino acid sequence is conserved in eukaryotes and is considered to be one of the widespread conserved "known unknown" eukaryotic genes [Gal10]. To unravel origin and molecular evolution of these genes in different eukaryotes, we carried out a comprehensive sequence, structural functional and phylogenetic analyses of BEM46 orthologs. During this study, we found that all eukaryotes have at least a single copy of a *bem46* ortholog. Upon scanning of these proteins from various spiecies, expansions leading into several paralogs in vertebrates and plants were identified. We illustrate insertion/deletions (indels) in the conserved domain of BEM46 protein, which allow differentiating fungal classes such as ascomycetes from basidiomycetes. Furthermore, we analyze several duplicates of this gene in different animal and plant genomes to understand possible mechanisms of evolution after separation from the fungal lineage. In addition, we unravel that BEM46 protein from *N. crassa* possess a novel endoplasmic-retention signal (PEKK) using GFP-fusion tagging experiments, hinting there is need to re-define the motifs in conserved in various protein sequences as over a million of genome sequences will be available in next decade.

References

[Mer09] Moritz Mercker, Krisztina Kollath-Leiß, Silke Allgaier, Nancy Weiland and Frank Kempken. The BEM46-like protein appears to be essential for hyphal development upon ascospore germination in Neurospora crassa and is targeted to the endoplasmic reticulum Curr Genet 55:151–161, 2009

[Gal10] Michael Y. Galperin and Eugene V. Koonin. From complete genome sequence to 'complete' understanding? Trends Biotechnol 28:398-406, 2010.

Establishment and analysis of fungal kinomes

Yousef Shbat, Abhishek Kumar and Frank Kempken

Abteilung für Botanik mit Schwerpunkt Genetik und Molekularbiologie,
Botanisches Institut, Christian-Albrechts-Universität zu Kiel, Germany.
akumar@bot.uni-kiel.de | fkempken@bot.uni-kiel.de

Many cellular processes are regulated by phosphorylation via protein kinases. To unravel the understanding of protein phosphorylation, genome-wide analysis of protein kinase complements (known as 'kinomes') is performed in eukaryotic species. About 2% of eukaryotic genes are protein kinases [Man02]. In our study we employed bioinformatics and comparative genomics to determine the fungal kinome in an evolutionary and functional context. Kinases are major regulators of cellular processes in fungi, in similar fashion as they regulate other eukaryotes. For example, 77 viable mutants for ser/thr kinase genes (of 86 in total) were identified in *N. crassa* and 57% illustrated at least one growth or developmental phenotype [Par11]. A study of 30 fungal genomes revealed that kinase number and extra domains play instrumental role in fungal kinome classification [Kos10]. Currently, there are more than 100 fungal genome sequences known. Hence, there is a need of more comprehensive analysis of fungal kinomes. We have established kinomes of about 80 fungi. Based on kinome size, fungi can be divided into two groups: (i) fungi with large kinome (average size 160), and (ii) fungi with small kinome (average size 60). We will illustrate how kinases in combination with about 20 other protein domains, evolved to perform their roles in different signaling cascades in these fungi. Furthermore, we annotate and analyse these kinases for evolutionary mechanisms operating in fungal kinomes.

References

[Man02] Manning *et al.* The Protein Kinase Complement of the Human Genome. Science, Vol. 298 no. 5600 pp. 1912-1934, 2002.

[Par11] Park *et al.* Global Analysis of Serine-Threonine Protein Kinase Genes in Neurospora crassa. Eukaryot Cell. 10(11): 1553–1564, 2011.

[Kos10] Kosti *et al.* Comparative analysis of fungal protein kinases and associated domains BMC Genomics, 11:133, 2010.

Classification by descent: Toward genetics-based taxonomy of RNA viruses

Chris Lauber[1], Igor A. Sidorov[1], Alexander A. Kravchenko[2], Dmitry V. Samborskiy[2], Andrey M. Leontovich[2], and Alexander E. Gorbalenya[1,2,3]

[1]Molecular Virology Laboratory, Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands
[2]A.N. Belozersky Inst. of Physico-Chemical Biology and [3]Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia

When a new virus is discovered a principal analysis step is its comparison, and thus classification, with known taxa. Virus classification is crucial for understanding the natural diversity of viruses and is commonly achieved by expert virologists who rely on a multitude of virus characteristics. For many newly described viruses, however, only the genome sequence is reported these days. To address this challenge, we recently developed a computational approach, named DEmARC, to classifying viruses of a large monophyletic group (family or order) using only genome data [1]. DEmARC is used to objectively devise the levels of a hierarchical classification as well as a threshold on pairwise genetic divergence below which two viruses are grouped together for each level.
Here we present DEmARC-mediated classification results for RNA viruses from three very different families: picornaviruses (positive-sense RNA genomes of ~7.5 kb), coronaviruses (positive-sense RNA, ~30 kb), and filoviruses (negative-sense RNA, ~19 kb); these three datasets also differ strongly in respect to virus sampling size which ranges from several dozens to hundreds. Each virus group contains major human pathogens including, respectively, poliovirus, SARS coronavirus, and ebolavirus. Generally, DEmARC results show excellent agreement to the respective taxonomy of the family (e.g. [2]). Additionally, it offers biological implications including few notable deviations from taxonomy that concern already classified viruses as well as the prediction of still unknown genetic diversity in the family [3]. We aim at making DEmARC-based classification generally available for the advancement of virus taxonomy in times when the genome sequence, the footprint of evolution, is the only information available for many viruses.

[1] Lauber, Gorbalenya (2012) *Journal of Virology* 86:3890
[2] Knowles, et al. (2012) Family *Picornaviridae*. In: King, et al. (Eds.) Virus taxonomy IX. *Academic Press*, pp 855-880
[3] Lauber, Gorbalenya (2012) *Journal of Virology* 86:3905

Mathematical modelling of oxygen diffusion: Does cellular oxygen consumption cause gradients that influence intracellular oxygen sensors?

Samantha Nolan, Oliver Sawodny, Michael Ederer
Institute for System Dynamics, University Stuttgart
samantha.nolan@isys.uni-stuttgart.de

Facultative anaerobic bacteria like *Escherichia coli* sense oxygen for control of gene expression to switch between aerobic and anaerobic metabolism. The global transcription factor FNR coordinating this switch contains a [4Fe-4S] cluster which reacts directly with oxygen in the cytoplasm.

During 100% aerobiosis, i.e. complete oxidation of glucose to carbon dioxide without by-product formation, the oxygen concentration in the medium is about 1 μM in the steady state of continuous culture, corresponding to 600 molecules per cell volume. However, the membrane bound oxidases consume up to 290 000 molecules per second during respiration [Ale00]. This high consumption in the membrane compared to the small number of molecules in the cytoplasm leads to the question whether large oxygen gradients occur near and within a cell. To address this question two mathematical models were established.

The first ordinary-differential equation model is a spatially lumped model showing the behaviour of oxygen in the three compartments medium, membrane and cytoplasm and its impact on FNR inactivation, hereby showing the effects of different oxygen concentrations, diffusion coefficients and reaction rates. The second model using partial-differential equations focuses on the oxygen gradients in consideration of the three-dimensional cell and environment. For the nominal literature-based parameter values, both models show only very small oxygen gradients suggesting that the terminal oxidases, FNR and the oxygen electrodes sense the same oxygen concentration. However, the values of the model parameters are partly uncertain. We analyze the model with respect to the uncertain parameters and characterize cases where gradients play a vital role.

References

[Ale00] S. Alexeeva et al., “Effects of limited aeration and of the ArcAB system on intermediary pyruvate catabolism in Escherichia coli.,” *Journal of bacteriology*, 182(17):4934-40, 2000.

Modelling the switch from type I to type II apoptosis during crosstalk of interleukin-1β and Fas ligand signalling in cultivated hepatocytes

Julia Sanwald[1]*, Anna Lutz[2]*, Mathias Könczöl[2], Oliver Sawodny[1], Irmgard Merfort[2], Michael Ederer[1]
1 Institute for System Dynamics, University of Stuttgart, Germany;
2 Department of Pharmaceutical Biology and Biotechnology, Albert Ludwigs University Freiburg, Freiburg, Germany;
** these authors contributed equally to the work*
julia.sanwald@isys.uni-stuttgart.de

Fas ligand, a prominent inducer of apoptosis, and the pro-inflammatory cytokine interleukin-1β, that is known to induce no cell death, are important signalling molecules in the natural environment of the liver. Experiments revealed a crosstalk effect when primary mouse hepatocytes cultured on collagen were stimulated with both IL-1β and FasL showing that IL-1β sensitizes cells to FasL-induced caspase-3/7 activation.

This sensitizing effect is due to a signalling switch from the type I pathway of apoptosis usually activated in cultivated primary hepatocytes upon FasL treatment to the type II pathway. Type II signalling was verified by Bid dependency and cytochrome c release. Although caspase-3 activity is usually considered as a suitable indicator for apoptosis, combined stimulation with IL-1β and FasL did not lead to increased cell death.

A dynamical model of the crosstalk of IL-1β and FasL signalling was developed supporting the mechanism of IL-1β-induced sensitization with respect to caspase-3/7 activation as well as a model-based hypothesis explaining the absence of the sensitizing effect with regard to cell viability. Model analysis revealed a strong dependency of caspase-3/7 activity on the protein levels of the Bcl-2 family and cFLIP suggesting large population heterogeneity. It was hypothesized that the population upon IL-1β/FasL treatment may divide into two subpopulations with one executing apoptosis via the type II pathway showing high caspase-3 activation and the other escaping from apoptosis showing almost no caspase-3 activity. Simulating such behaviour by varying initial conditions of anti-apoptotic proteins according to a normal distribution demonstrates the possibility of a heterogeneous cell population to exhibit similar viability but increased caspase-3/7 activity if cytochrome c release occurs very late during time course of the experiment. Therefore, considering hetero-

geneity within a cell population as well as time points of stimulation and measurement are of high importance for explaining the experimentally observed effects.

Discovery of emphysema/COPD-relevant molecular networks from an A/J mouse COPD inhalation study by means of Reverse Engineering and Forward Simulation **(REFS™)**

Yang Xiang[1]*, Ulrike Kogel[1], Stephan Gebel[2], Michael J. Peck[1], Manuel C. Peitsch[1], Viatcheslav R. Akmaev[3], Boris Hayete[3], Jignesh Parikh[3], John Caprice[3], Julia Hoeng[1], Iya Khalil[3]

[1]*Philip Morris Research and Development, CH-2000 Neuchâtel, Switzerland*

[2]*Philip Morris Research Laboratories GmbH, D-51149, Germany.*
[3]*GNS Healthcare, Cambridge, Massachusetts, USA.*
Yang.Xiang@pmi.com

Chronic Obstructive Pulmonary Disease (COPD) is a pulmonary disease characterized by progressive airflow limitation caused by parenchyma destruction (emphysema) and small airway disease (obstructive bronchiolitis). It is associated with an abnormal inflammatory response of the lung to noxious particles and gases. The factors that determine an individual heavy smoker's risk for developing COPD have not yet been defined, and in all likelihood COPD is the result of the combined actions of genetic and environmental factors.

Emphysema is a component of COPD. To investigate the mechanism leading to the development of emphysema in A/J mice exposed to two doses of cigarette smoke (CS) for various periods of time, we analyzed different biological endpoints (MMP-9, MMP activity, TIMP-1 and lung weight) together with whole gene expression profiles obtained from lung tissue. A novel and powerful method, known as reverse engineering and forward simulation (REFS™) [Xing et al., 2011], was then employed to generate an ensemble of molecular networks. Simulations showed that this ensemble method could successfully recover the measured experimental data. By simulating thousands of *in silico* gene knockdown experiments with this ensemble of networks we were able to identify thirty-three molecular key

drivers that may have an effect on the level of at least one of the biological endpoints measured. After functional analysis, it appears that those key drivers are involved in inflammation processes and lung parenchyma destruction.

References

Xing, H., et al., Causal modelling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. PLoS Comput Biol, 7(3): p. e1001105, 2011.

BiSQuID: Bisulfite Sequencing Quantification and Identification

Cassandra Falckenhayn, Günter Raddatz and Frank Lyko
German Cancer Research Center
g.raddatz@dkfz.de

Bisulfite sequencing (BS) has developed as a standard method to analyze epigenetic methylation patterns. Over the last years applications for high throughput BS have become more and more important. Besides whole genome bisulfite sequencing which is mainly carried out using Illumina platforms, 454 pyrosequencing has been established as a standard method for BS-sequencing of short amplicons. Although a number of programs and web services have been introduced to enable users to process their BS-sequences [Roh10][Kum08], all of these services have different shortcomings and limitations, especially they are not actively maintained to keep up with the ongoing development of BS-sequencing. Here we present BiSQuID (Bisulfite Sequencing Quantification and Identification), a tool which is designed to process 454 BS-sequences in an easy and convenient way. BiSQuID includes a database, which stores the definition of primers, barcodes etc., making it possible to keep track of series of experiments. BiSQuID includes algorithms to either process long amplicons containing mismatches/indels as well as bigger amounts of short reads without indels and automatically adjusts the appropriate processing options. BiSQuID visualizes the results using heatmaps and saves the results as comprehensive packages which can be easily stored and exchanged between researchers.

References

[Roh10] Rohde, C., Zhang, Y., Reinhardt, R. and Jeltsch, A.
BISMA - Fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences.
BMC Bioinformatics 2010, 11:230

[Kum08] Kumaki, Y., Oda, M. & Okano, M.
QUMA: quantification tool for methylation analysis
Nucleic Acids Res. 36, 2008, W170-W175

Large–scale organization of metabolic network models

Jens Einloft, Jörg Ackermann, Joachim Nöthen, Ina Koch
Institute of Computer Science, Molecular Bioinformatics, Goethe University Frankfurt
Einloft@bioinformatik.uni-frankfurt.de

The topological properties of metabolites in networks have been studied extensively in the literature to address the large–scale cellular organization of organisms [JTA+00, RSM+02]. Structurally, the networks are characterized by, e.g., the distribution of node degree, the distribution of cluster coefficient, and distribution of shortest path length. To improve the statistical significance of the results an average process over a collection of networks is commonly applied. The structural properties of metabolites are rather well characterized, but the structural properties of reactions have been ignored so far. We analyze topological properties of 1846 whole genome metabolism models published recently in the database Path2Models (http://www.ebi.ac.uk/biomodels-main/path2models). The models are matched to bipartite graphs with directed edges. In this way information about the direction of substance flow and the role of reactions is transferred to the graphical representation. An appropriate mathematical formalism is given by the theory of Petri nets. We present various structural properties of metabolites and reactions for genome metabolism models. The models contains up to 6644 metabolites and up to 9950 reactions. The ratio of reactions and metabolites is linearly correlated with an average of 1.66 reactions per metabolite. The existence of superhubs is a characteristic feature of the networks. The highest vertex degree of metabolites goes up to 2237 and the highest vertex degree of reactions is 53. The distribution of the vertex degree and the cluster coefficient of metabolites describe a long memory correlsation with typical Hurst parameters in the range of 2.2 to 2.9.

References

[JTA+00] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[RSM+02] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabsi. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297(5586):1551–1555, 2002.

Atlas of gene-specific transcription factors and their epistatic relationships in yeast

Katrin Sameith[1], Marian Groot Koerkamp[1], Dik van Leenen[1], Mariel Brok[1], Tineke Lenstra[1], Joris Benschop[1], Sander van Hooff[1], Berend Snel[2], Patrick Kemmeren[1], Frank Holstege[1]
[1]*Molecular Cancer Research, University Medical Centre Utrecht, NL*
[2]*Department of Biology, Utrecht University, NL*
k.sameith@umcutrecht.nl

Transcription plays a key role in cellular processes and its regulation is of paramount importance. For understanding transcription on a systems level, it is important to discern target genes of gene-specific transcription factors (GSTFs). Here, we analyzed 183 deletion mutants of GSTFs in *Saccharomyces cerevisiae* by DNA microarrays. Determination of functional target genes by comparing gene expression changes to published binding data reveals that repression of transcription plays a much more prominent role than previously anticipated. Of all surveyed GSTFs that could be categorized into activators and/or repressors, activators account for less than 54%. The remaining 46% of GSTFs are repressors (37%) or have a dual function (9%). The unanticipated high number of gene-specific repressors indicates that in the yeast *S. cerevisiae*, chromatin is not as restrictive to transcription as previously thought and indicates that a considerable part of the gene-specific machinery is aimed at restricting unwanted transcription. The systematic analysis of individual GSTF deletion mutants also shows that many GSTFs can be removed without affecting gene expression. This can partly be explained by redundancy relationships, whereby the loss of one GSTF can be compensated by the presence of another GSTF. To systematically investigate epistatic relationships between GSTFs such as redundancy, we additionally generated gene expression profiles for 73 GSTF double deletion mutants. For each GSTF pair, epistatic effects are determined by comparing gene expression changes in the two single mutants and their respective double mutant. Redundancy relationships in particular are found both between GSTFs with common binding targets and GSTFs with intimately related functions. The results demonstrate that gene expression profiling of double deletion mutants facilitates detailed understanding of combinatorial control through GSTF pairs.

GC content dependency of Open Reading Frame prediction

Martin Pohl[1], Günter Theißen[2] and Stefan Schuster[1]
[1]*Department of Bioinformatics,* [2]*Department of Genetics, Friedrich Schiller University Jena*
m.pohl@uni-jena.de

A frequently used approach for detecting potential coding regions is to search for stop codons. In random or non-coding DNA one can expect every 21st trinucleotide to have the same sequence as a stop codon since 3 out of 64 trinucleotides are stop codons in the standard genetic code. In contrast, the open reading frames (ORFs) of most protein-coding genes are considerably longer. Thus, the stop codon frequency in coding sequences deviates from the background frequency of the corresponding trinucleotides [CO02]. This has been utilized for gene prediction, in particular, in detecting protein-coding ORFs. Traditional methods based on stop codon frequency are based on the assumption that the GC content is about 50 %. However, many genomes show significant deviations from that value. With the presented method we can describe the effects of GC content on the selection of appropriate length thresholds of potentially coding ORFs. Thus, we can derive the maximum GC content for which ORF length is practicable as a feature for gene prediction methods and the resulting false positive rates. We demonstrate the feasibility of this method by applying it to the genomes of the bacteria *Rickettsia prowazekii*, *Escherichia coli* and *Caulobacter crescentus*. Usually, several methods for gene finding need to be combined. Interestingly, for genomes with low GC content such as that of *R. prowazekii*, the presented method provides remarkably good results even when applied alone[PS12].

References

[CO02] Carpena, P., Bernaola-Galvan, P., Roman-Roldan, R. and Oliver, J.L. A simple and species-independent coding measure. *Gene*, 300:97-104, 2002.

[PS12] Pohl, M., Theißen, G. and Schuster, S. GC content dependency of Open Reading Frame prediction via stop codon frequencies. *Gene*, submitted, 2012.

MetaProteomeAnalyzer: A software tool specifically developed for the functional and taxonomic characterization of metaproteome data

T. Muth[1*], R. Heyer[2], A. Behne[1], F.Kohrs[2], D. Benndorf[2], E. Rapp[1] and U. Reichl[1,2]

1. Max Planck Institute for Dynamics of Complex Technical Systems

2. Otto-von-Guericke University Magdeburg, Chair of Bioprocess Engineering

muth@mpi-magdeburg.mpg.de

The functional analysis of highly complex microbial communities is a rather challenging discipline in proteomics. However, there is no alternative to its use if biological measures are required to improve the economic efficiency of biogas plants and waste water treatment plants, to remediate contaminated soil and water or to investigate the complex interactions in the human gut. Currently, analysis and interpretation of data derived from LC-MS/MS experiments present a major bottleneck in metaproteomics. In contrast to pure-culture proteomics, metaproteome samples are heterogeneous and much more complex. Moreover, for the major part of the microorganisms, protein sequence information is not available which results in a low protein identification rate. To overcome limits of existing solutions, we developed a software tool for the functional and taxonomic characterization of metaproteomics data we called MetaProteomeAnalyzer (MPA). Beside an advanced protein identification by a combination of multiple database search algorithms (Crux, OMSSA, X!Tandem, Inspect), and a spectral library search, the identification of proteins from unsequenced species by *de novo* sequencing and a BLAST search are included in the workflow. In addition, features for the analysis of the taxonomic diversity of a microbial community and for the functional classification of proteins are included.MPA constitutes a protein identification platform specialized on metaproteomics, which facilitates the analysis of complex microbial communities and increases significantly the number of identified proteins.

Structural Insights into the Inhibition of GSK-3β by 1-amino-2,4,5-trihydroxy-7-methyl-anthracene-9,10-diones

Katja S. Lerche[1], Doris Mahn[1], Robert Günther[2], Hans-Jörg Hofmann[2] and Rolf Gebhardt[1]

[1]Institute of Biochemistry, Faculty of Medicine, University of Leipzig, [2]Institute of Biochemistry, Faculty of Biosciences, Pharmacy and Psychology, University of Leipzig

Katja.Lerche@medizin.uni-leipzig.de

GSK-3β is an important regulator in insulin signalling. An increased activity is closely associated with diabetes type II and obesity. But GSK-3β and CK-2 are also key targets in Wnt/β-catenin signalling and cancer. This double-edged functionality makes it necessary, to find inhibitors with potential anti-diabetic qualities, without a simultaneous up-regulation of the main Wnt-target β-catenin. For this inhibitor screening, docking studies were performed. These studies were operated with the genetic algorithm of AutoDock 3.05. The evaluation of 50 independent docking runs, employing a Lamarckian genetic algorithm, were initiated with randomly chosen target simulations and iterated through energy evaluations. In a panel of more than 30 anthraquinones, 4-aminoethylamino-emodin was identified as a potent and selective GSK-3β inhibitor with an IC_{50}-value in the nano molar range, whereas structural analogous without an aminoethylamino side chain were selective inhibitors of CK-2. On the one hand, a sterical hindrance existed for CK-2 through the amino-alkyl residue. On the other hand, this amino-alkyl-amino-residue fitted perfectly the acidic-hydrophobic-acidic (Asp133-Tyr134-Val135-Pro136-Glu137) amino acid arrangement of GSK-3β, the crucial step for the inhibitory potency. In particular, 4-aminoethylamino-emodin exhibited astonishing features in and outside the insulin signalling pathway.

References

Gebhardt R, Lerche KS, Götschel F, Günther R, Kolander J, Teich L, Zellmer S, Hofmann HJ, Eger K, Hecht A, Gaunitz F. 4-Aminoethylamino-emodin-a novel potent inhibitor of GSK-3beta-acts as an insulin-sensitizer avoiding downstream effects of activated beta-catenin. J Cell Mol Med, 14(6A): 1276-93, Jun 2010.

Theoretical study of two minus mating type specific dehydrogenases of the zygomycete *Mucor mucedo*

Sabrina Ellenberger[a], Stefan Schuster[b], Johannes Wöstemeyer[a]
[a] *Chair of General Microbiology and Microbial Genetics, Friedrich-Schiller-University Jena, 07743 Jena, Neugasse 24, Germany*
[b] *Department of Bioinformatics, Friedrich-Schiller-University Jena, 07743 Jena, Ernst-Abbe-Platz 2, Germany*
Sabrina.Ellenberger@uni-jena.de

In the fungal group zygomycetes trisporic acid is one of the major fields of research. Its isomers and derivatives mediate sexual and parasitic communication. This trisporoids act as pheromones and are produced by cooperative synthesis between a plus and a minus mating type. 4-dihydromethyltrisporate dehydrogenase (TSP1, [1]) and 4-dihydrotrisporin dehydrogenase (TSP2, [2]) are two minus mating type specific dehydrogenases of this biosynthetic pathway. We made predictions about secondary structures and binding pockets with 3DLigandSite [3] based on sequences from *Mucor mucedo* for the two dehydrogenases. TSP1 is an aldo-keto reductase with a TIM-barrel structure. TSP2 belongs to the short-chain dehydrogenases and is characterized by a Rossmann fold. Although both are NADP-dependent oxidoreductases, acting on trisporoids, the protein structures and ligand-binding interactions of them are different. We analyzed the localization of NADP and trisporoid ligands within this proteins with the help of docking studies and searched for regions which affect substrate specificity. So TSP1 has two possible regions for trisporoid-binding. TSP2 in contrast has just one region although the protein surfaces of TSP1 and TSP2 look similar around the active site. That means, there is one open area on the protein surface, which allows a wide range of conformations and a narrow tunnel.

References

[1] Katrin Czempinski, Volker Kruft, Johannes Wöstemeyer, and Anke Burmester. 4-Dihydromethyltrisporate dehydrogenase from *Mucor mucedo*, an enzyme of the sexual hormone pathway: purification, and cloning of the corresponding gene. *Microbiol.*, 142:2647–2654, 1996.

[2] Jana Wetzel, Olaf Scheibner, Anke Burmester, Christine Schimek, and Johannes Wöstemeyer. 4-Dihydrotrisporin-Dehydrogenase, an Enzyme of the Sex Hormone Pathway of *Mucor mucedo*: Purification, Cloning of the corresponding Gene, and Developmental Expression. *Eukaryotic Cell*, 8(1):88–95, 2009.

[3] Mark N. Wass, Lawrence A. Kelley, and Michael J. E. Sternberg. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, 38:W469–W473, 2010.

Root-Games between Plants: Predicting Tendency for Cooperation along environmental Gradients

Sebastian Germerodt[1], Jana Schleicher[1], Katrin Meyer[2], David Ward[3], Stefan Schuster[1], Kerstin Wiegand[2]

[1] *Department of Bioinformatics, University Jena*

[2] *Department of Ecosystem Modelling, University Göttingen*

[3] *School of Life Sciences, University of KwaZulu-Natal*

Sebastian.Germerodt@uni-jena.de

The degree of root overlap between neighboring plant individuals is determined by the net-outcome of plant interactions. We developed a model based on Evolutionary Game Theory and an individual-based model (IBM) to demonstrate how a positive (facilitative) process (i.e. hydraulic lift) may change the root proliferation strategy between plants competing for belowground resources and how this influences the degree of root overlap. In our approaches we interpreted facilitative interaction as a 'cooperation' strategy. In the game approach, the simplified situation of two plants competing for the same region of soil is represented. The calculation of the net payoffs includes a surplus (due to facilitative processes) if plant roots overlap, costs of root development, and competition for nutrients. We show that the payoff for cooperation is not constant and the evolutionarily stable strategies change with availability of water and nutrients. This implies that the tendency for root overlap may also change along environmental gradients. In the IBM approach, we evaluated how the tendency to cooperate, and thus the degree of root overlap, depends on local resource availability, the strength of the facilitative process, and the costs for root development. Under low nutrient conditions and an average to high water availability aggregations of plants with a high tendency for defection emerged in the model. At low water availability, plants with a high tendency for cooperation accumulated. In areas with high availability of resources, a mixed pattern of the two strategies emerged. In sum, environmental gradients can change the net outcome of interactions, such as cooperation and competition, and can establish local patterns of interaction modes and varying degrees of root system overlap.

A Spatio-Temporal Modeling Framework to Simulate Host-Pathogen Interactions

Johannes Pollmächer and Marc Thilo Figge
Applied Systems Biology, Leibniz-Institute for Natural Product Research and Infection Biology, Hans Knöll Institute (HKI), Friedrich Schiller University Jena, Jena, Germany
johannes.pollmaecher@hki-jena.de

The growing field of Image-based Systems Biology with its increasing amount of spatio-temporal analyses requests for adequate modeling instances to drive the cycle of Systems Biology [HHK+12].

We developed a hybrid spatio-temporal modeling framework that allows performing simulations at the cellular and molecular scale. Our special focus is on cell-migration and cellular interactions based on input parameters that are obtained from the image-analyses of experimental microscopy studies for host-pathogen interactions. To reduce the time-complexity for the process of interaction-detection we applied a grid-based monitoring-system to keep track of each cellular position in continuous space with time.

In a continuous space representation, a brute force approach scales proportional to the square of the number of interacting cells. The present approach makes use of a method based on neighbor lists. That is particularly appropriate for biological systems since cells interact locally with each other. As a consequence, the simulation time in the present approach remains close to linear in the number of interacting cells. Furthermore, the method of neighbor lists is advantageous over the Delaunay-triangulation method, since the latter needs significantly more computational effort regarding the management of the cellular agents.

We apply the modeling framework to phagocyte-fungus interactions, where the phagocyte is able to uptake and kill the fungus, while the fungus has the ability to escape the immune response by killing the phagocyte through hyphae formation and piercing of the phagocyte. The spatio-temporal modeling framework is adaptable to a wide range of multi-cellular systems.

References

[HHK+12] Fabian Horn, Thorsten Heinekamp, Olaf Kniemeyer, Johannes Pollmächer, Vito Valiante, and Axel A Brakhage. Systems biology of fungal infection. *Frontiers in Microbiology*, 3(00108), 2012.

Single cell track analysis of two-photon microscopy on T_h17 cells in the gut

Zeinab Mokhtari and Marc Thilo Figge
Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology Hans-Knöll-Institute (HKI), Friedrich Schiller University Jena, Jena, Germany
zeinab.mokhtari@hki-jena.de

Two-photon microscopy is a modern imaging technique for tracking single cells *in vivo* in order to analyze the migration and interaction behavior of cells under physiological conditions. Often, statistical analysis of these data are performed where mean quantities are computed from averaging over all cell tracks. As a consequence, all information about the cellular positions in the biological sample is averaged out. We show that important information in the imaging data can get lost in this way and that a single cell track analysis is required that takes cellular positions into account to correctly interpret the migration behavior of cells.

We analyze two-photon microscopy data on T helper 17 (T_h17) cells from *in vivo* imaging in the gut of mice. These cells produce interleukin 17 (IL17) and are involved in the immune response against bacteria and fungi. Based on cell tracking data and the visualization of blood vessels in microvilli of the murine gut, we first perform a statistical analysis and arrive at the conclusion that the migration behavior of T_h17 cells resembles a random walk. However, a single cell track analysis reveals that the migration behavior is only seemingly random.

The single cell track analysis involves a reconstruction of the dynamic blood vessel from the three-dimensional imaging data (z-stacks) and allows to determine the distance of individual T_h17 cells in each time point relative to the moving surface of the blood vessel. This dynamic analysis of single cell tracks shows that T_h17 cells are in fact not performing a random walk but migrate in close proximity to the moving surface of blood vessels. In the most general sense, this investigation demonstrates the importance of single cell track analysis over statistical analysis for interpreting two-photon microscopy data.

PAA - A New R Package for Autoimmune Biomarker Discovery with Protein Microarrays

Michael Turewicz, Maike Ahrens, Caroline May, Helmut E. Meyer and Martin Eisenacher
Medizinisches Proteom-Center, Ruhr-University Bochum, Bochum, Germany
michael.turewicz@rub.de

Background: Protein microarrays like the ProtoArray (Life Technologies, Carlsbad, CA, USA) are used for autoimmune antibody screening studies to discover biomarker panels. For ProtoArray data analysis the software Prospector (Life Technologies) is often used because it provides an advantageous feature ranking approach ("M score", [Lov07]). Unfortunately, Prospector provides no capabilities regarding multivariate feature selection, classification, batch effect adjustment and computational biomarker candidate validation.
Results: Therefore, we have adopted Prospector's M score approach and implemented a new R package called Protein Array Analyzer (PAA) that provides these features and a complete data analysis pipeline for ProtoArray and other single color microarray data that come in gpr file format. After optional data pre-processing and M score-based feature preselection a multivariate feature selection is performed. For this purpose, a backwards elimination (wrapper) approach ("gene shaving" with random forest, [JDC+04]) has been implemented. For the selection and validation of stable panels a frequency-based approach has been adopted ([BTC09]). Furthermore, different plots and results files can be obtained to outline the analysis results.
Conclusions: We propose the new R package PAA for protein microarray data analysis. PAA has been used to successfully analyse several different ProtoArray data sets (e.g. "Parkinson", "Alzheimer", "Amyotrophic Lateral Sclerosis"). Thereby, its suitability for biomarker discovery with protein microarrays has been shown.

References

[BTC09] S. Baek, C. A. Tsai, and J. J. Chen. Development of biomarker classifiers from high-dimensional data. *Brief Bioinform*, 10(5):537–46, 2009.

[JDC+04] H. Jiang, Y. Deng, H. S. Chen, L. Tao, Q. Sha, J. Chen, C. J. Tsai, and S. Zhang. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, 5(81), 2004.

[Lov07] B. Love. *The Analysis of Protein Arrays*, pages 381–402. CRC Press, 2007.

The role of α-ketoglutarate dehydrogenase in stabilizing the flux through the citric acid cycle

Dorothee Girbig and Joachim Selbig
Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany
Institute of Biochemistry and Biology, University of Potsdam, Germany
girbig@mpimp-golm.mpg.de

The citric acid cycle (TCA cycle) plays a central role in energy metabolism and in providing precursors for biosynthetic pathways. Consequently, this pathway relies on mechanisms that guarantee a stable flux through the cycle despite the perturbations occurring at the branch points to adjacent pathways. Here we investigate the intrinsic stabilizing mechanisms of the neuronal TCA cycle by structural kinetic modeling (SKM) [SGSB06]. SKM enables the systematic assessment of stabilizing sites in metabolic networks by deriving a parameterized representation of the systems Jacobian matrix, in which the model parameters encode information about individual enzyme-metabolite interactions. The parameter space can be analyzed by decision trees in order to detect stabilizing patterns [GGS12]. In agreement with previous experimental and kinetic modeling results [BBH12], we find that the enzyme α-ketoglutarate dehydrogenase (AKGDH) plays an important role in controlling the flux through the cycle. We investigate how the AKGDH reaction has to be coordinated with other TCA cycle enzymes in order to ensure stable flux, and how these patterns behave under different experimental conditions.

References

[BBH12] Nikolaus Berndt, Sascha Bulik, and Hermann-Georg Holzhütter. Kinetic Modeling of the Mitochondrial Energy Metabolism of Neuronal Cells: The Impact of Reduced α-Ketoglutarate Dehydrogenase Activities on ATP Production and Generation of Reactive Oxygen Species. *International Journal of Cell Biology*, pages 1–11, 2012.

[GGS12] Dorothee Girbig, Sergio Grimbs, and Joachim Selbig. Systematic analysis of stability patterns in plant primary metabolism. *PLoS ONE*, 7(4):e34686, 2012.

[SGSB06] Ralf Steuer, Thilo Gross, Joachim Selbig, and Bernd Blasius. Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences*, 103(32):11868–11873, 2006.

Incorporating Proteome Similarities for Improved Species Abundance Estimation in Metaproteomics

Anke Penzlin, Martin S. Lindner, Bernhard Y. Renard
Research Group Bioinformatics (NG 4), Robert Koch-Institute, Nordufer 20, 13353 Berlin, Germany
RenardB@rki.de

One goal in metaproteomics is the quantitative taxonomic assessment of microbial community compositions in environmental samples by mass spectrometry based proteome analysis. In particular, relative quantification of taxons opens doors for diagnostics or microbial community comparison. When mapping mass spectra from a metaproteomic experiment to a proteome database, the aim is to find the correct assignment for each peptide to a species contained in the database. In a naïve approach, taxons can be quantified using peptide match counts for each proteome, but there are severe problems discerning species with highly similar proteome sequences due to ambiguous matches.

Here, we present a method to estimate true proteome abundances via peptide identification by using reference proteome similarities in a non-negative LASSO approach [RKS+08, LR12]. The proteome similarities are estimated by mapping simulated spectra derived from the proteome of a single organism to all proteomes. We simulate the spectra with MSSimulator [BAAR11]. Our approach produces precise abundance estimates of similar species: We demonstrate on simulated data that our approach is able to improve on the naïve approach and reduces error by more than 65%.

References

[BAAR11] C. Bielow, S. Aiche, S. Andreotti, and K. Reinert. MSSimulator: Simulation of Mass Spectrometry Data. *Journal of Proteome Research*, 10(7):2922–2929, 2011.

[LR12] M.S Lindner and B.Y. Renard. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Research*, doi:10.1093/nar/gks803, (in press) 2012.

[RKS+08] B.Y. Renard, M. Kirchner, H. Steen, J.A.J. Steen, and F.A. Hamprecht. NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9(1):355, 2008.

Parameter Estimation by Simulated Annealing for Models of Whole-Blood Infection Assays with Candida Albicans

Teresa Lehnert, Marc Thilo Figge
Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute (HKI), Friedrich-Schiller-University Jena, Jena, Germany
teresa.lehnert@hki-jena.de

Mathematical modeling of biological systems requires the estimation of a priori unknown parameter values. The precise estimation of model parameters reveals insight into the relative importance of individual processes in complex biological systems. We simulate time-resolved data obtained from human whole-blood infection assays with *Candida albicans* by numerically solving a mathematical model in terms of coupled ordinary differential equations. The optimal set of model parameters is obtained from a self-written algorithm that performs simulated annealing based on the Metropolis Monte Carlo scheme.

The algorithm searches for the global minimum in the deviations of the simulation results relative to the experimental data by randomly exploring the space of parameter sets. Whether or not a new set of parameters is accepted depends on the weighted least square error (wLSE) between the experimental and simulated data points. The annealing procedure involves a continuous decrease of the initially high acceptance of parameter sets with higher wLSE during successive Metropolis Monte Carlo steps. Two different procedures of error evaluation have been implemented: In the *individual procedure*, the wLSE for each differential equation is compared with the experimental data separately, while in the *joint procedure* the overall wLSE is chosen for the comparison.

We applied this approach to estimate the parameters of mathematical models on time-resolved data obtained from human whole-blood infection assays with *C. albicans*. The immune defense against this human-pathogenic fungi can be realized by different routes in parallel, e.g. involving the complement system, cellular interactions with granulocytes and monocytes, as well as antimicrobial factors. The mathematical models aim at elucidating the relative importance of these routes to fend off *C. albicans* in human blood.

Semi-Automated Evaluation of Microbial Observables from High-Throughput Time-Lapse Microscopy

Stefan Helfrich, Alexander Grünberger, Dietrich Kohlheyer,
Wolfgang Wiechert, Katharina Nöh
IBG-1: Biotechnology, JARA-HPC, Forschungszentrum Jülich GmbH
s.helfrich@fz-juelich.de

Image-based single cell analysis is a vital approach to make time-resolved observations and gain insights into the development of microbial, isogenic populations. State-of-the-art lab-on-a-chip devices allow for the long-term cultivation of various bacteria under well-controlled conditions in a highly parallel manner. The manual generation and annotation of lineage trees from collected time-lapse images is a time-consuming and error-prone task that becomes infeasible for high-throughput experimentation. Thus, computational tools are needed to support the large-scale analysis and statistical evaluation of experiments.

Image analysis is particularly challenging for low contrast images resulting from high cell densities and volatile mitosis events. We describe a semi-automated image analysis pipeline that addresses these challenges. The computational toolset is capable to (i) pre-process raw images, (ii) identify cells, (iii) separate cell clusters, (iv) track cells throughout an image sequence, (v) visualize microbial lineage trees derived from tracking-results, and (vi) annotate the trees with additional information. The image analysis pipeline bases on advanced, available methods, which are adapted for our specific requirements and implemented using the ImageJ platform [SRE12]. Additionally, a novel Python-based visualization tool for lineage trees is presented along with a custom file format to enrich the trees with additional information, e.g., fluorescence.

The analysis pipeline is demonstrated on data collected in a growth study of *Gluconobacter oxydans*. Time-dependent morphometric information, extracted from the image sequences, can be used to estimate observables of microbial populations, e.g., growth rates and cell lengths. This opens perspectives for the comparison of any two populations and their characterization with regard to heterogeneity. With that, a tool for the statistical analysis of high-throughput growth experiments is available.

References

[SRE12] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7):671–5, June 2012.

Analysis of RNA-Seq data after knockdown of *amer* gene family members in zebrafish

Stefan Pietsch, Birgit Perner, Christoph Englert
Leibniz Institute for Age Research - Fritz Lipmann Institute, Jena
spietsch@fli-leibniz.de

August 17, 2012

Mutations in the X-linked tumor suppressor gene *Wtx*, also called *amer1*, cause Wilms tumor and sclerosing bone dysplasia. A function for *amer1* in the Wnt Signaling pathway has been suggested, however, not jet proven [1, 2]. The role of *amer1* and its gene family members *amer2* and *amer3*, especially in the Wnt signaling pathway, should be investigated in a loss-of-function experiment followed by RNA-Seq. Morpholino knockdown of these genes in fertilized zebrafish eggs had been performed and RNA was isolated from two different embryonic stages. Subsequently, the RNA samples were sequenced using Illumina technology.

The aim of this study was to analyse the sequenced RNA fragments (reads) and to search for differentially expressed genes (DEG) after *amer1/2/3* knockdown. Therefore the reads were aligned with the genome (Bowtie, TopHat) and gene expression levels were calculated. Different methods were used to find significant DEGs, e.g. DESeq, edgeR and Cufflinks.

A comparison of various ways to analyze RNA-Seq data was made and the results were used to further investigate the potential functions of selected DEGs. As a result, *amer1* and *amer2* seem to have similar functions, which are different from *amer3*. The role of *amer1* as a negative regulator of the Wnt signaling pathway could be confirmed. New potential targets for the amer family members were found.

References

[1] Major, M. B. & et al. (2007). Wilms tumor suppressor WTX negatively regulates WNT/beta-catenin signaling. *Science (New York, N.Y.)*, 316(5827), 1043–6.

[2] Tanneberger, K. & et al. (2011). Amer1/WTX couples Wnt-induced formation of PtdIns(4,5)P2 to LRP6 phosphorylation. *The EMBO journal*, 30(8), 1433–43.

Transcriptomic analysis of the adult life stage of the invasive Colorado potato beetle (Leptinotarsa decemlineata) using Roche 454

Abhishek Kumar[1,2] & Alessandro Grapputo[1]

[1] Dept. of Biology University of Padova. Padova, Italy
[2]Abteilung für Botanische Genetik und Molekularbiologie, Botanisches Institut und Botanischer Garten, Christian-Albrechts-Universität zu Kiel, Kiel, Germany.
akumar@bot.uni-kiel.de | grapputo@bio.unipd.it

The Colorado potato beetle (*Leptinotarsa decemlineata*) is a major pest and a serious threat to potato cultivation throughout the northern hemisphere. Despite its high importance for invasion biology, phenology and pest management, little is known about *L. decemlineata* from a genomic perspective and only larval trancriptome data is reported. Hence, we subjected adult European *L. decemlineata* transcriptome samples to 454-FLX massively-parallel DNA sequencing to characterize a basal set of genes from this species. Our focus was in diapause-specific genes and genes involved in pesticide and *Bacillus thuringiensis* (Bt) resistance. Using 454-FLX pyrosequencing, we obtained a total of 395,877 reads with 23,952,311 bp, which assembled into 66,675 high quality expressed sequence tags (48,942 contigs and 17,733 singletons). We established repository of genes of interest such as those for diapause, with 89 out of 106 diapause-specific genes described in *Drosophila montana*, and insecticide resistance, including 150 cytochrome P450 monoxygenases (CYPs), 43 Glutathione S-transferases (GSTs), 4 catalases (CAT), 15 superoxide dismutases (SOD), 22 glutathione peroxidases (GPX) and 67 esterases. We found 37 putative miRNAs and we predicted a significant number of single nucleotide polymorphisms (2,281) and microsatellite loci (484).

Furthermore, we unravelled 16 sequences similar to *Wolbachia* endosymbiont indicating possible infection of the species with this endosymbiont bacteria. We also assembled and provided basic annotation for the publically larval Roche 454 reads and a combined adult and larval dataset and made available for community usage. This report of the assembly and annotation of the transcriptome of adult *L. decemlineata* offers new insights into diapause-associated and insecticide-resistance-associated genes in this species and provide a foundation for comparative studies with other species of insects. The data will also open new avenues for researchers using *L. decemlineata* as a model species, and for pest management research.

Deep roots and stepwise evolutionary history of the vertebrate head sensory systems

Martin Sebastijan Šestak[1], Vedran Božičević[2], Robert Bakarić[1,3], Vedran Dunjko[4], Tomislav Domazet-Lošo[1]

[1]*Ruđer Bošković Institute, Zagreb, Croatia*
[2]*Evolutionsbiologie, Ludwig-Maximilians-Universität München-Biozentrum, Planegg-Martinsried, Germany*
[3]*Max-Planck-Institut für Evolutionsbiologie, Plön, Germany*
[4]*School of Informatics, University of Edinburgh, UK*
msestak@irb.hr

The complex head sensory organs are traditionally regarded as the innovations of the vertebrate lineage. Emergence of the sensory systems is connected to their developmental and evolutionary origins from the cranial placodes and the neural crest, which are thought to have emerged simultaneously at the base of the vertebrates [Nor05]. Other studies, however, point to the deeper evolutionary roots and stepwise emergence of these vertebrate features during the chordate-vertebrate transition [BL12]. As systematic studies that could resolve these controversies are currently lacking, the evolutionary origin of the presumed vertebrate sensory innovations is still unclear. To give a fresh perspective on the adaptive history of the head sensory organs in the ancestors of vertebrates, we applied a phylostratigraphic approach [DLT10] on the large scale *in situ* expression data of the developing zebrafish *Danio rerio*. Opposite to the traditional predictions, we find that dominant adaptive signals in the analyzed sensory structures are mostly preceding the evolutionary advent of vertebrates. The visual system shows the earliest adaptive signals that correspond to the bilaterian-chordate transition. The majority of the head sensory organs, including the olfactory, otic and lateral line systems have leading adaptive signals at the origin of Olfactores, ancestors of tunicates and vertebrates. The only structures that qualify as genuine vertebrate innovations include trigeminal ganglion, adenohypophysis, and neural crest derivatives, which is in line with the traditional expectation. Taken together these results reveal deep and stepwise adaptive history of the vertebrate sensory innovations.

References

[BL12] Bronner ME, LeDouarin NM. Development and evolution of the neural crest: An overview. *Developmental Biology* 366:2–9, 2012.

[DLT10] Domazet-Loso T, Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468:815–818, 2010.

[Nor05] Northcutt RG. The new head hypothesis revisited. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 304B:274–297, 2005.

Coupled Mutation Finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations

Mehmet Gültas[1], Martin Haubrock[1], Nesrin Tüysüz[2], and Stephan Waack[1]
[1] *University of Göttingen, Germany,* [2] *Erasmus Medical Center, Rotterdam, The Netherlands*

Multiple sequence alignments (MSAs) of homologous protein sequences characterize compensatory mutations between non-conserved residue sites that are crucial for the protein stability and functionality. These residue sites are as important as the strictly conserved positions for the understanding of the structural basis of protein functions and for the identification of functionally important residue positions. However, the detection of significant compensatory mutation signals in MSAs is often complicated by noise. A challenging problem in bioinformatics is the separation of significant signals between two or more non-conserved residue sites from the phylogenetic noise and unrelated pair signals. In order to determine significant compensatory mutation signals, we developed the Coupled Mutation Finder (CMF) [GHTW]. The CMF combines two models: i) an MSA-specific statistical model of significant signals based on multiple testing procedures; ii) a novel entropy-based metric to upscale BLOSUM62 dissimilar compensatory mutations. While the former quantifies the error made in terms of the false discovery rate, the latter shows how dissimilar compensatory mutations have affected genomic sequences in the course of evolution. To demonstrate the performance and functionality of the CMF, we analyzed the structurally or functionally important positions of two human proteins, namely epidermal growth factor receptor (EGFR) and glucokinase (GCK). The CMF detects in these two proteins disease associated amino acid mutations (non-synonymous single nucleotide polymorphisms (nsSNPs)), not strictly conserved catalytic or binding sites, and residue positions that are nearby one of these sites which are likely to affect protein stability or functionality. Our results suggest that the CMF is a helpful tool to predict and to investigate compensatory mutation sites of structural or functional importance in proteins. The CMF server is freely accessible at http://cmf.bioinf.med.uni-goettingen.de.

References

[GHTW] M. Gültas, M. Haubrock, N. Tüysüz, and S. Waack. Coupled Mutation Finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations. *Under Review.*

Automated Image Analysis of Hodgkin lymphoma

Alexander Schmitz, Hendrik Schäfer, Tim Schäfer, Jörg Ackermann, Norbert Dichter, Sylvia Hartmann, Martin-Leo Hansmann, Ina Koch
Goethe-University Frankfurt am Main, Germany
ina.koch@bioinformatik.uni-frankfurt.de

Hodgkin lymphoma is an unusual type of lymphoma, arising from malignant B-cells. Morphological and immunohistochemical features of malignant cells and their distribution differ from other cancer types. Based on systematic tissue image analysis, computer-aided exploration can provide new insights into Hodgkin lymphoma pathology.

In our poster, we report results from an image analysis of CD30 immunostained Hodgkin lymphoma tissue section images. To the best of our knowledge, this is the first systematic application of image analysis to a set of tissue sections of Hodgkin lymphoma. We have implemented an automatic procedure to handle and explore image data in Aperio's SVS format. We use pre-processing approaches on a down-scaled layer to separate the image objects from the background. Then, we apply a supervised classification method to assign pixels to predefined classes. Our pre-processing method is able to separate the tissue content of images from the image background. We analyzed three immunohistologically defined groups, non-lymphoma and the two most common forms of Hodgkin lymphoma, nodular sclerosis and mixed type. We found that nodular sclerosis and non-lymphoma images exhibit different amounts of CD30 stain, whereas mixed type exhibits a large variance and overlaps with the other groups.

The results can be seen as a first step to computationally identify tumor regions in the images. This allows us to focus on these regions when performing computationally expensive tasks like object detection in the high-resolution layer. We apply a CellProfiler pipeline to detect primary objects. Therefore the image is split into its color stains using a color deconvolution approach. By setting a treshold in the CD30 stain image we identify CD30 positive objects and calculate their shape descriptors. We classify the cells regarding the size, elongation and compactness. We present results for a small set of nodular sclerosis, mixed type and non-lymphoma images.

DDIS - A new algorithm for comparing gene interaction graphs

Vindi Jurinovic and Ulrich Mansmann
Institute for Medical Bioinformatics, Biometry and Epidemiology - IBE, LMU Munich

jurinovic@ibe.med.uni-muenchen.de

1 Introduction

Recently, interest in gene interaction networks has led to an increase in the number of algorithms for network estimation. Comparing gene networks between two conditions is an important issue, since it can discover differences where differential expression analysis fails. Usually, differences between networks are assessed by first estimating the networks, and then comparing their edges in terms of true and false positives. However, this can introduce more errors into the analysis, since edges in the network already bear uncertainties from the estimating process.

2 Method

We propose a new algorithm that compares the networks indirectly and therefore does not require their estimation. The algorithm is based on the comparison of path plots produced by Lasso regression [Tib96]. For each node, we introduce a statistic that tests if its direct neighborhood differs between two conditions. Finally, a resampling procedure is used to test specific null-hypotheses on differential correlation structures of two graphs. It is also possible to apply the test on a set of nodes or a whole graph.

3 Results

In a simulation study with random and scale free graphs, the latter representing biological networks, we show that our approach has reasonable statistical properties. We compare it to a similar proposal from the literature [G$^+$10] and show that our approach offers several advantages. Furthermore, we apply the test on a real data set from a study of acute lymphoblastic leukemia in children, and compare our findings with published results [KS04].

References

[G$^+$10] Ryan Gill et al. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11:59, 2010.

[KS04] Dennis Kostka and Rainer Spang. Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20:194–199, 2004.

[Tib96] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1), 1996.

Unraveling stress resistance from metagenomic sequence of Socompa stromatolites

Kurth, D.[1], Albarracin, V.H. [1,2], Revale, S.[3], Rascovan, N.[3], Timmermann, B. [4], Vazquez, M.[3], Farias, M.E.[1]

[1]*Planta Piloto de Procesos Industriales y Microbiológicos (PROIMI), CCT, CONICET. San Miguel de Tucumán, Argentina*
[2]*Max-Planck-Institute for Bioinorganic Chemistry, Mülheim-an-der-Ruhr, Germany*
[3]*Instituto de Agrobiotecnologia Rosario (INDEAR), Rosario, Argentina*
[4]*Max-Planck-Institute for Molecular Genetics, Berlin, Germany.*

dgkurt@gmail.com

High Altitude Andean Lakes (HAAL) are extreme environments exposed to high ultraviolet radiation (UVR), elevated salinity, and high concentration of arsenic, located in northwestern Argentina 3500 meters above sea level. Living stromatolites were recently identified by our group in one of these lakes, Laguna Socompa [AD11]. These ancient microbial communities and its location in these extreme environments are key to understand microbial processes related to early life on Earth, and the possibility of life in Mars.

In this work a metagenomic approach was applied to study Socompa stromatolites. Environmental sequencing of total DNA obtained from Socompa stromatolites allowed a phylogenetic characterization and functional analysis of the community. Further analysis of metagenomic sequences allowed the identification of genes related to two processes of outmost importance for the survival of bacteria in the HAAL: resistance to UV radiation and arsenic tolerance. High UV irradiation on these extreme environments enhances the selection of systems to withstand this radiation. It can be expected that the HAAL photolyases have unique properties allowing maximum activity and improved resistance to UV [AP12]. Analysis of genes related to arsenic resistance showed the presence of unusual extrusion pumps.

These are only isolated examples of the rich genetic biodiversity found in these harsh habitats. Further studies will reveal new genes with interesting biotechnological applications such as production of metabolites with potential therapeutic effects, as antibiotics, or UV resistant compounds.

References

[AD11] Albarracín, V.H., J.R. Dib, *et al*. A harsh life of indigenous proteobacteria at the andean mountains: Microbial diversity and resistance mechanisms towards extreme conditions. In Proteobacteria: Phylogeny, metabolic diversity and ecological effects. M. L. Sezenna, Nova Publishers, 2011.

[AP12] Albarracín, V.H., G.P. Pathak, *et al*. Extremophilic Acinetobacter Strains from High-Altitude Lakes in Argentinean Puna: Remarkable UV-B Resistance and Efficient DNA Damage Repair. Orig Life Evol Biosph 42(2-3): 201-221, 2012.

Theoretical study of lipid accumulation in the liver – Implications for nonalcoholic fatty liver disease

J. Schleicher[1], R. Guthke[2], H.G. Holzhütter[3] and S. Schuster[1]
1 *Department of Bioinformatics, University Jena*
2 *Hans-Knöll-Institute, Jena*
3 *Medical Department (Charité), University Berlin*
jana.foerster@uni-jena.de

A hallmark of the nonalcoholic fatty liver disease (NAFLD) is the accumulation of lipids in the liver (steatosis). We developed a simple mathematical model of the lipid dynamics in hepatocytes. Our spatially homogeneous model involves fatty acid intake, lipid degradation (oxidation) and triglyceride export. It takes into account that storage of triacylglycerol within hepatocytes leads to an enlargement of cell size [McC04]. Then, the swelling of hepatocytes reduces the cross-section of sinusoids and impairs hepatic microcirculation [Kon10]. Thus the supply with oxygen is reduced, which, in turn, impairs lipid degradation. The analysis of our model revealed a bistable behavior (two stable steady states) of the system. The first stable state is characterized by intact lipid degradation and only a low amount of stored lipids. This state may correspond to the healthy hepatocyte. The second stable state in our model is marked by a high amount of stored lipids and reduced lipid degradation caused by cell enlargement and impaired hepatic microcirculation. This state may correspond to the steatotic hepatocyte. An interesting outcome of our model is its reversibility. The system can switch from the steatotic state back to the healthy state by a reduction of fatty acid intake. This corresponds to the observation that changes of lifestyle of NAFLD patients, especially reduced caloric diet, can cure steatosis.

References

[McC04] R. McCuskey, Y. Ito, G.R. Robertson, M.K. McCuskey, M. Perry, and G.C. Farrell. Hepatic microvascular dysfunction during evolution of dietary steatohepatitis in mice. *Hepatology*, 40(2): 386-393, 2004.

[Kon10] K. Kondo, T. Sugioka, K. Tsukada, M. Aizawa, M. Takizawa et al. Fenofibrate, a Peroxisome Proliferator-Activated Receptor α Agonist, Improves Hepatic Microcirculatory Patency and Oxygen Availability in a High-Fat-Diet-Induced Fatty Liver in Mice. *Advances in experimental Medicine and Biology*, 662: 77-82, 2010.

Topology separation of discriminative sequence motifs located in membrane proteins with domains of unknown functions

Steffen Grunert, Florian Heinke and Dirk Labudde
Department of Mathematics, Natural and Computer Sciences,
University of Applied Sciences Mittweida
Technikumplatz 17, D-09648 Mittweida
steffen.grunert@hs-mittweida.de, florian.heinke@hs-mittweida.de,
dirk.labudde@hs-mittweida.de

Motivation: Membrane proteins play essential roles in cellular processes. Nutrient transport, signal and energy transduction or ion flow are only a few of the numerous functions enabled by membrane proteins. The analysis of membrane proteins has shown to be an important part in the understanding of complex biological processes in the context of proteomics and genomics. Genome-wide investigations of membrane proteins have revealed a large number of short, distinct sequence motifs. These motifs support the understanding of the features that are important for establishing stability and functionality of the folded membrane protein in the membrane environment [Gerst02]. Thus, membrane protein sequence motif analysis can be helpful in a number of applications, e.g. the investigation of mutant proteins and potential effects of mutagens.
Material & Methods: In the focus of our statistical analysis, 50 protein sequence motifs, reported in [Gerst02], were derived, identified and analyzed in 32 membrane protein families, including proteins with domains of unknown function (DUF). The dataset was derived from Pfam database [Punta12].
Results: The analysis of 32 DUF-families led to a novel approach which describes the separation of motifs by residue-specific distributions. Based on these distributions we can predict the topology state of the majority motifs in hypothesized membrane proteins with unknown topology.
Conclusion: It can be hypothesized that motifs with high prediction accuracy are essential structure-forming elements in membrane proteins, whereas other motifs might be family-specific elements for defining protein function.

References

[Gerst02] Y. Liu, D. M. Engelman, and M. Gerstein. Genomic analysis of membrane protein families: abundance and conserved motifs. Genome Biol., 3(10):research0054, Sep 2002.

[Punta12] Marco Punta, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L L. Sonnhammer, Sean R. Eddy, Alex Bateman and Robert D. Finn. The pfam protein families database. Nucleic Acids Res, 40(Database issue):D290–D301, Jan 2012.

Automated Encoding of Gene Regulatory Networks from Inference Tools in SBML

Bianca Hoffmann[1], Sebastian Vlaic[1], Andreas Dräger[2]

[1] *Leibniz Institute for Natural Product Research and Infection Biology Hans-Knöll-Institute, Beutenbergstr. 11a, D-07745 Jena, Germany*
[2] *Center for Bioinformatics Tuebingen (ZBIT), University of Tuebingen, D-72076 Tübingen, Germany*

{bianca.hoffmann, sebastian.vlaic}@hki-jena.de
andreas.draeger@uni-tuebingen.de

The understanding of the complex interaction processes within biological networks is the driving force of research in systems biology. One part of this endeavor is the regulation of gene expression, which gives information about the behavior and function of biological systems. To understand the complex mechanisms, mathematical models are used to reconstruct gene regulatory networks (GRNs). Since high-throughput technologies offer evermore data in this field, the number of models and their complexity permanently rises. With regard to interoperability and further development the use of a standard format for model representation is necessary. The Systems Biology Markup Language (SBML) [HFS$^+$03] covers this aspect and defines a standardized format for electronic model representation of mainly quantitative models describing biological processes at the chemical reaction level. Since GRNs do not mimic the cellular processes in such detail, automated encoding of GRNs in SBML in compliance with the specifications of the standard is a mandatory step towards interoperability, exchangeability and proper storing. We developed GRN2SBML, an interface which is based on the Java^{TM} library JSBML [DRD$^+$11], that provides methods for creating, reading and manipulating SBML documents. Considered model architectures are systems of linear ordinary differential equations. During the export, nodes of the network and the model itself can be annotated with diverse metadata such as Gene Ontology [ABB$^+$00] terms or references to publications, which improves the semantic content of information making model interpretation and analysis easier. All together GRN2SBML provides an easy way to encode GRNs in a standardized format. Instead of using specialized analysis software, respective to the original format of model representation, a wide range of SBML capable software tools can be applied to simulate and analyze the exported models.

References

[ABB+00] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[DRD+11] A. Dräger, N. Rodriguez, M. Dumousseau, A. Dörr, C. Wrzodek, N. Le Novère, A. Zell, and M. Hucka. JSBML: a flexible Java library for working with SBML. *Bioinformatics*, 27(15):2167–2168, 2011.

[HFS+03] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

Introducing Tree Topology Profiling for Meta-Analysis of Whole-Genome Phylogenies

Thomas Meinel[+] and Antje Krause[*]
[+]*Charite - University Medicine Berlin, Institute for Physiology, Structural Bioinformatics Group, Thielallee 71, 14195 Berlin, Germany*
[*]*FH Bingen, Bioinformatics, Berlinstr. 109, 55411 Bingen, Germany*
[+]Corresponding author: thomas.meinel@charite.de

A large number of whole-genome phylogenies have been inferred during the last decades to reconstruct the "Tree of Life". Underlying data models range from gene content of species to gene families and multiple sequence alignments of concatenated protein sequences. Diversity in data models together with the use of different tree reconstruction techniques, biological effects like horizontal gene transfer and the steadily increasing number of genomes have led to a huge diversity in published phylogenies [Hou09].

We introduce "tree topology profiling" as a new method to compare already published whole-genome phylogenies [MK12]. Particular topology alternatives found for an ordered list of bacterial clans reveal a topology profile that represents the analyzed phylogeny.

For an in-depth analysis of 47 selected published phylogenies we generate an additional set of seven phylogenies using different inference techniques and the SYSTERS-PhyloMatrix data model [MKL[+]05]. After tree topology profiling we separate artefactual from biologically meaningful phylogenies and associate particular inference results (phylogenies) with inference methodologies (inference techniques as well as data models).

References

[Hou09] C.H. House. *Horizontal Gene Transfer: Genomes in Flux*, volume 532, chapter The Tree of Life Viewed Through the Contents of Genomes, pages 141–161. Humana Press, 2009.

[MK12] T. Meinel and A. Krause. Meta-Analysis of General Bacterial Subclades in Whole-Genome Phylogenies Using Tree Topology Profiling. *Evolutionary Bioinformatics*, 8:489–525, 2012.

[MKL[+]05] T. Meinel, A. Krause, H. Luz, M. Vingron, and E. Staub. The SYSTERS Protein Family Database in 2005. *Nucleic Acids Res.*, 33(Database issue):D226–D229, 2005.

Comparative transcriptomics of *Arabidopsis thaliana* and *Arabidopsis lyrata*

Yvonne Pöschl[1], Carolin Delker[2], Jana Gentkow[2], Marcel Quint[2], and Ivo Grosse[1]
[1]*Institute of Computer Science, Martin Luther University Halle–Wittenberg, Halle (Saale)*
[2]*Leibniz Institute of Plant Biochemistry, Halle (Saale)*
poeschl@informatik.uni-halle.de

Microarrays are still widely used for expression studies and are a cheap alternative to deep-sequencing platforms. Affymetrix microarrays are designed for a *target species*, where genes are represented by probe sets consisting in average of 11 probes. The hybridization signals of all probes of a probe set are summarized to one value representing the expression of the corresponding gene. However, microarrays are typically not appropriate for comparative transcriptomics studies that require other *query species* due to differences in their transcripts. This problem has been addressed by [Ham05] who hybridize genomic DNA (gDNA) of the query species to the microarray designed for the target species. As removing improper probes lead to more accurate gene expression values, they exclude probes with low gDNA hybridization signals from further analyses. Here, we propose an alternative approach that does not require the intermediate wet-lab step but relies on in-silico sequence comparison alone. We apply both approaches to *Arabidopsis thaliana* and its sister species *A. lyrata* and find that the gDNA approach eliminates 58% of the probes of the ATH1 microarray, while the sequence-based approach eliminates 55%. We assess the quality of the two approaches by comparing the expression values of 29 genes to those obtained by qRT-PCR. We find that the log fold-changes resulting from the sequence-based approach show a higher Pearson correlation to the log fold-changes of the qRT-PCR control experiments ($c = 0.954$) than the log fold-changes resulting from the gDNA approach ($c = 0.898$). This suggests that using sequence information is not only cheaper and less time consuming, but can also be more accurate for array-based comparative transcriptomics, than hybridizing gDNA.

References

[Ham05] Hammond, J.P. et al. Using genomic DNA-based probe-selection to improve the sensitivity of high-density oligonucleotide arrays when applied to heterologouse species. *Plant Methods*, 1(10), 2005.

Alignment of flowgrams to strings

Marcel Martin
Bioinformatics for High-throughput Technologies, TU Dortmund
marcel.martin@tu-dortmund.de

A read from the 454 and Ion Torrent sequencers is natively represented as a *flowgram*, which is a sequence of pairs of a nucleotide and its (fractional) intensity. Conversion of flowgrams to strings incurs a loss of information, which can be avoided by aligning the reads in their flowgram form to a reference. Fractional intensities and ambiguous run lengths are then resolved at alignment time by choosing the most likely read sequence given both the nucleotide intensities and the reference sequence.

Previous methods either convert the reference string to an (artificial) flowgram [VJZL08] or convert the flowgram to a string while retaining intensity information [LAP11]. Only the second method supports editing events on the nucleotide level (insertions, deletions and substitutions), through an extension of the Smith-Waterman algorithm, but certain events cannot be modelled.

We introduce the concept of a *string-flowgram alignment*, which is an intuitive way to describe how sequencing errors due to incorrect intensities and editing events add up to result in the observed flowgram. It is simply a list of nucleotide/intensity tuples or gaps that are each paired with the substring of the reference that they align to. A string-flowgram alignment can, in particular, represent substitutions within and also insertions between homopolymer runs. We give an efficient, dynamic programmig algorithm that finds an optimal string-flowgram alignment, which is more general than previous methods as it does not need to convert between representations.

References

[LAP11] F Lysholm, B Andersson, and B Persson. FAAST: Flow-space Assisted Alignment Search Tool. *BMC Bioinformatics*, 12:293, 2011.

[VJZL08] V Vacic, H Jin, J Zhu, and S Lonardi. A probabilistic method for small RNA flowgram matching. *Pac Symp Biocomput*, pages 75–86, 2008.

RNA structure: Does secondary structure define a fold ?

Nikolai Hecker and Andrew E. Torda
Zentrum für Bioinformatik, Universität Hamburg
hecker@zbh.uni-hamburg.de

RNA has become remarkably fashionable in the last few years and new functions are attributed to it on a weekly basis. Thinking on its structure is, however, still dominated by the idea that most features can be explained by loops and helices. The helices, in turn, reflect the pattern of Watson-Crick base-pairing. If one believes that RNA folds hierarchically (helices and base-pairs form first), then it would seem that secondary structure prediction is the first step towards prediction of full 3-dimensional coordinates. With this study, we try to see to what extent this is useful. If one knew the secondary structure of an RNA molecule with no errors, would this define reasonable 3-D coordinates ? This was tested by taking a set of known 3D coordinates, extracting the base pairs and using these as restraints in distance geometry calculations. This leads to large numbers of conformations which are consistent with the correct base-pairing. We examined the spread of allowed conformations and then tried clustering these, using an objective measure (the Davies-Bouldin index) to investigate the structure of the allowed conformational space. This leads to a disappointing conclusion. Even if one could predict secondary structure with absolutely no errors, one would still be far from real tertiary structure prediction.

eProS - A Database and Toolbox for large-scale Analyses of energetic Properties that determine Protein Structure and Function

Florian Heinke, Daniel Stockmann, Stefan Schildbach and Dirk Labudde
Department of Mathematics, Natural and Computer Sciences, University of Applied Sciences Mittweida
forename.surname@hs-mittweida.de

Gaining information about structural and functional features of newly identified proteins is often a difficult task. This information is crucial for understanding sequence-structure-function relationships of target proteins and, thus, essential in comprehending the mechanisms and dynamics of the molecular systems of interest.
Using protein energy profiles is a novel approach that can contribute in addressing such problems. An energy profile corresponds to the sequence of energy values which are derived from a coarse-grained energy model. Energy profiles can be computed from protein structures or predicted from sequences. As shown, correspondences and dissimilarities in energy profiles can be applied for investigations of protein mechanics and dynamics [HL12, MMK07].
We developed eProS (energy profile suite), a database which provides about 76,000 pre-calculated energy profiles as well as a toolbox for addressing numerous problems of structure biology. Energy profiles can be browsed, visualised, calculated from uploaded structure or predicted from sequence. Furthermore, it is possible to align energy profiles of interest or compare energy profiles with the entire eProS database to identify significantly similar energy profiles and, thus, possibly relevant structural and functional relationships. Additionally, annotations and cross-links retrieved from numerous databases (i.e. Gene Ontology, SCOP, CATH) provide a broad view of potential biological correspondences. eProS is freely available at http://bioservices.hs-mittweida.de/Epros/.

References

[HL12] F. Heinke and D. Labudde. Membrane protein stability analyses by means of protein energy profiles in case of nephrogenic diabetes insipidus. *Comput Math Methods Med*, 2012:790281, 2012.

[MMK07] D. Mrozek, B. Malysiak, and S. Kozielski. An Optimal Alignment of Proteins Energy Characteristics with Crisp and Fuzzy Similarity Awards. In *FUZZ-IEEE'07*, pages 1–6, 2007.

Network-based Prioritization and Functional Characterization of Disease Genes

Nadezhda T. Doncheva[1,*], Tim Kacprowski[2], Mario Albrecht[1,2]
[1]*Max Planck Institute for Informatics, Saarbrücken, Germany*
[2]*University Medicine Greifswald, Greifswald, Germany*
[*]*E-mail: nadezhda.doncheva@mpi-inf.mpg.de*

The field of candidate disease gene prioritization has evolved fast in the last years [DKA12]. Many methods have been developed, some of which make use of functional annotations or integrate multiple data sources using statistical learning techniques. Additionally, network-based approaches that exploit large-scale interaction data have become an indispensable tool in this field. However, it is still difficult to achieve high prioritization performance in all cases of medical research. Here, we highlight key points for further methodological improvements of disease gene prioritization.

Phenotypes can strongly differ in their genetic characteristics as well as in the amount of research dedicated to them. Therefore, prioritization methods should be tailored to specific phenotypes or groups of phenotypes. This might be accomplished by carefully selecting the appropriate sources of biomedical knowledge and ranking strategies for candidate genes. Both ideas together can be combined into novel network-based approaches. In a recent disease case study, we integrated protein interaction data and functional similarities based on the Gene Ontology to generate disease-specific networks for inflammatory bowel diseases. By analyzing and comparing these networks, we revealed the functional overlap of similar inflammatory phenotypes. Furthermore, we developed a new Cytoscape plugin and adapted different topological measures to rank candidate genes in molecular interaction networks. We also applied suitable rank aggregation algorithms to merge the resulting rankings.

Reference:

[DKA12] Nadezhda T. Doncheva, Tim Kacprowski, and Mario Albrecht. Recent approaches to the prioritization of candidate disease genes. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(5):429-442, 2012.

Emerging new dynamic behavior from the coupling of subsystems: the case of EGFR trafficking and signaling

Carolina Gallo López [1,2], Lars Kaderali [1]

[1] *Institute for Medical Informatics and Biometry (IMB), Dresden University of Technology, Germany.*

[2]*University of Evry-Val-d'Essonne, France.*

carolina.gallolopez@ens.univ-evry.fr, lars.kaderali@tu-dresden.de

Endocytosis is a major mechanism of receptor signal attenuation allowing their removal from the cell surface. Deregulation of Epidermal Growth Factor Receptor (EGFR) signaling components have been shown to have a critical role in the development and progression of many types of cancer. We quantitatively investigated how EGFR trafficking dynamics controls EGFR downstream signaling pathways as a function of EGF concentration. Two ODE-based mathematical models describing different but complementary processes were coupled into an integrated model. The first model describes EGF dose-dependent EGFR trafficking by two routes: clathrin-mediated endocytosis (CME) and clathrin-independent endocytosis (CIE) [3]. The second model describes two EGFR signaling components: SOS recruitment (i.e. Ras activation) and PLCγ phosphorylation [2]. Model integration resulted in an increase in the number of components and reactions. At low ligand concentrations, the receptor is almost exclusively internalized through CME, which leads to receptor recycling and sustained signaling, with limited degradation. At higher concentrations, the receptor is internalized through the two endocytic routes, CIE leads instead to an ubiquitin-dependent degradative pathway, limiting both the duration and the intensity of receptor-dependent transient signaling response. Following the signaling endosome hypothesis [1], the differential phosphorylated receptor distribution in cellular compartments allows differentiating spatiotemporal signaling component activities, as well as, adding signaling specificity. Parametric sensitivity revealed a dependency on the targeted signaling components (e.g. SOS recruitment and PLCγ phosphorylation) and on the EGF concentrations, for instance the

recycling rate of CIE-internalized complex and receptor and the receptor deubiquitination rate in endosome turned out to be critical parameters for SOS recruitment and PLCγ phosphorylation at high EGF concentration.

References

[1] Howe, Valletta, Rusnak, Mobley. *Neuron* 32: 801–814, 2001.
[2] Kholodenko, Demin, Moehren, Hoek. *J.Biol.Chem.* 274, 30169-81, 1999.
[3] Suryavanshi. Mathematical model to uncover the role of receptor ubiquitination in dose dependent EGFR trafficking. *Ph.D. thesis*. University of Heidelberg. 2012.

Prediction of MicroRNAs in a human fungal pathogen

Janine Freitag[1], Jörg Linde[1], Ronny Martin[2], Oliver Kurzai[2], Reinhard Guthke[1], Dominic Rose[3]
[1]*Research Group Systems Biology / Bioinformatics - Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Jena, Germany.* [2]*Research Group Fungal Septomics - Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Jena, Germany.* [3]*Chair of Bioinformatics, Institute of Computer Science, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany*
{janine.freitag, joerg.linde, ronny.martin, oliver.kurzai, reinhard.guthke}@hki-jena.de, dominic@informatik.uni-freiburg.de

MicroRNAs are small, non-coding RNAs which regulate gene expression through target mRNA binding and translational repression. They are well studied in plants, insects and mammalia but little is known in fungi, although microRNA-like RNAs were found in *Neurospora crassa* [ea10]. Additionally, Dicer- and Argonaute-Proteins, required for miRNA preprocessing, were detected in *Candida albicans* [ea09], the major invasive fungal pathogen of humans. This commensal fungus mainly infects immunocompromised individuals, resulting in high mortality rates. One important virulence trait is the switch between its commensal and pathogenic form going hand in hand with the change from yeast to hyphae. Even though the transcriptional network covering this switch has been extensively studied, it can not completely explain all observed yeast hyphae transitions. For this reason, we hypothesise that microRNAs play a role in this regulatory network.
The morphologic forms are closely linked to different pH values. To this end, we performed Next Generation Sequencing of small RNAs when *Candida albicans* cells are grown on pH 4 and 8, respectively. We mapped the reads to the *Candida albicans* genome and identified non protein coding regions, which are highly expressed. We examined these regions and used a cluster method based on LocaRNA [ea07] to compare the sequence and the secondary structure to arbitrarily selected known microRNAs from *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Homo sapiens*.
We found several sequences that form the characteristical hairpin structures of microRNAs. Furthermore some of the known microRNAs were

clustered contiguously to the highly expressed regions of the *Candida albicans* genome meaning that they closely resemble the known microRNAs. A few of them are also differently expressed in both datasets, indicating a role in the yeast hyphae switch. We will further analyse these sequences in the laboratory.

References

[ea07] Will et al. *PLoS Comput Biol*, 3(4):e65, 2007.

[ea09] Drinnenberg et al. *Science*, 326(5952):544–550, 2009.

[ea10] Lee et al. *Molecular cell*, 38(6):803–814, 06 2010.

Next-Newtomics: The next generation repository for bioinformatical interpreted ht-omics data from the newt Notophthalmus viridescens

Marc Bruckskotten, Jens Preussner, Thilo Borchardt, Mario Looso and Thomas Braun
Max-Planck-Institute for Heart and Lung Research, Ludwigstrasse 43, 61231 Bad Nauheim, Germany

Notophthalmus viridescens, an urodelian amphibian, represents an excellent model-organism to have insights into molecular processes driving regeneration. These achievements have been severely hindered by paucity and poor annotation of coding nucleotide sequences.
Comprehensive repositories for standard model organisms provide access to all levels of molecular biological data. For non-standard model organisms only little information from publically accessible is available.
In 2012, the Newt-omics repository database [Bru12], based on an EST dataset from the newt is extended by a sequential hybrid de novo assembly strategy of NGS data from different sources (454, Illumina, Sanger) [2]. This extends the existing sequence database by 120922 non-redundant transcripts (N50=975bp). The annotation-routines identified 38384 new putative annotateable transcripts. Newt-omics also extends the set of peptides identified by mass spectrometry, which was used to validate 15000 transcripts as protein coding. A deeper analysis unveiled 826 proteins specific only for urodeles [Loo12].
Next Newt-omics covers almost the entire transcriptome of the newt and represents a wealthy data-warehouse providing bio-molecular information on transcriptome and proteome level as well as functional characterization and experimental data on transcriptome and proteome level. The experimental data comprises transcript level analysis of time-series during heart and lens regeneration and cartilage.
The “Next Newt-omics” database is freely available online without registration at http://newt-omics.mpi-bn.mpg.de.

References

[Bru12] Bruckskotten M, Looso M, Reinhardt R, Braun T, Borchardt T.
Newt-omics: a comprehensive repository for omics data from the newt Notophthalmus viridescens.

Nucleic Acids Res. 2012 Jan;40(Database issue)
[Loo12] Looso M,Preussner J, Sousounis K, Bruckskotten M, Michel C, Reinhardt R, Höffner S, Krüger M, Tsonis P, Borchardt T and Braun T
Proteomics assisted de novo assembly of the newt transcriptome identifies new protein families expressed during tissue regeneration,Submitted to Genome Research, 2012

Machine Learning on Physiological Parameters to Perform Mouse Strain Characterization.

Mark Moeller and Georg Fuellen
Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, Germany
mark.moeller@uni-rostock.de

Mice are among the most widely used animals for research studies. Being genetically well defined, characterized and having an manageable life expectancy, they are well useful for longevity experiments.

Our aim has been to find the best physiological parameters to allow the successful allocation of a sample mouse to the correct strain. Therefore we used RandomForest™ as a machine learning approach and variable importance as the feature selection algorithm.

The MousePhenomeDatabase by the Jackson laboratory provided us with valuable longevity-related data records for different experiments including several blood analyses of an average of 30 inbred strains of mice. This time dependent data (6 months, 12M, 18M/20M, 24M) allowed us to check how the predictive quality of the implied model is changing with the increasing age of the mice.

It can be concluded that a complete blood count provides very good data to differentiate between laboratory mouse strains with an accuracy of over 80%. The best results have been achieved using the more detailed peripheral blood leucocyte profiles with accuracies over 90%, supporting the strain specificity of the immune system.

The accuracy of prediction is age dependent. The younger the mouse, the better are the results. This might be explained by the increasing variance of the measured values for older mice.

Validation of a metabolic model for *Arabidopsis thaliana*

Joachim Nöthen[1], Enrico Schlei ff[2], Jens Einloft[1], Jörg Ackermann[1], Ina Koch[1]
[1]*Molecular Bioinformatics, Institute of Computer Science, Johann-Wolfgang-Goethe Universität, Frankfurt*
[2]*Molecular Cell Biology of Plants, Institute of Molecular Plant Science, Johann-Wolfgang-Goethe Universität, Frankfurt*
joachim.noethen@bioinformatik.uni-frankfurt.de

In Systems Biology, *Arabidopsis thaliana* is a model organism for plants. Interrelations of metabolites can be represented in metabolic networks and be further analyzed with the help of various modeling techniques. Petri nets are a useful method which already has successfully been applied for pathway analysis of metabolic network for other plants, e.g. for sucrose breakdown in potato tuber [KJH05].
Validation is a crucial point in the modeling process. Petri net models can be validated based on their structural properties [VHK03]. Yet, computation of t-invariants requires exponential space, and is not feasible for large networks. An essential property for metabolic networks is the CTI property, i.e., each transition takes part in at least one t-invariant [ZOS03].
We created a metabolic Petri Net for *Arabidopsis thaliana*, consisting of 134 compounds connected via 242 reactions. To validate the model, we prove the scale-free property. Previous publications show that metabolic networks are scale-free [JTA$^+$00]. We apply network reductions to our network, to show that our network is CTI. The reduced network consists of 47 compounds connected by 104 reactions.

References

[JTA$^+$00] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[KJH05] Ina Koch, Björn H. Junker, and Monika Heiner. Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, 21(7):1219–1226, 2005.

[VHK03] Klaus Voss, Monika Heiner, and Ina Koch. Steady state analysis of metabolic pathways using Petri nets. *In Silico Biology*, 3(3):367–387, 2003.

[ZOS03] Ionela Zevedei-Oancea and Stefan Schuster. Topological Analysis of Metabolic Networks Based on Petri Net Theory. *In Silico Biology*, 3:323–345, 2003.

Functional Module Discovery in Molecular Interaction Networks using ModuleGraph

Tim Kacprowski[1,*], Sarah Foerster[2], Elke Hammer[3], Uwe Völker[3], Christoph A. Ritter[2], Mario Albrecht[1]

[1]*Department of Bioinformatics, Institute of Biometrics and Medical Informatics, University Medicine Greifswald, Germany*

[2]*Clinical Pharmacy, Institute of Pharmacy, Ernst Moritz Arndt University of Greifswald, Germany*

[3]*Department of Functional Genomics, Interfaculty Institute of Genetics and Functional Genomics, University Medicine Greifswald, Germany*

[*]*E-mail: tim.kacprowski@uni-greifswald.de*

Molecular interaction networks consist of many pairwise interactions of molecules. The interacting molecules have to be grouped by function to gain biological and medical insights into healthy and diseased states. To this end, the incorporation of functional modules such as protein complexes or pathways into interaction networks adds important functional information.

Therefore, we developed ModuleGraph, a new plugin for Cytoscape. It facilitates the discovery and visualization of functional modules in networks. To highlight the presence and composition of modules and their connections in a network, modules are represented as additional nodes linked to the module members. For example, protein complexes are connected to their constituent proteins. If modules share one or more constituents, they are interconnected. ModuleGraph also computes two important scores for each module. The *module representation score* indicates how many elements of a module are contained in the network. The *module connectivity score* quantifies how well these elements are connected to each other in the network.

In an application study, we used ModuleGraph to explore the interactome of the epidermal growth factor receptor (EGFR) for the analysis of functionally related protein complexes and signaling pathways. EGFR plays a major role in cancer and regulates proliferation, growth, and differentiation. Protein complexes isolated from EGF-stimulated or unstimulated cells were characterized by proteomics techniques using state-of-the-art mass spectrometry. The detected protein associations were then visualized using ModuleGraph and compared with known protein interactions and complexes.

Finding approximate gene clusters with Gecko 2

Sascha Winter[1], Katharina Jahn[2], Leon Kuchenbecker[2,3,4], Jens Stoye[2] & Sebastian Böcker[1]
[1] *Chair for Bioinformatics, Friedrich-Schiller-University Jena*
[2] *Genome Informatics, Faculty of Technology, Bielefeld University*
[3] *International Max Planck Research School for Computational Biology and Scientific Computing, Berlin*
[4] *Berlin-Brandenburg Center for Regenerative Therapies, Charité-Universitätsmedizin Berlin, Berlin*
sascha.winter@uni-jena.de

Gene-order based comparison of multiple genomes provides signals for functional analysis, and the rapid increase in sequenced genomes necessitates bioinformatics tools for finding gene clusters in hundreds of genomes. We present Gecko 2, a software for finding gene clusters in 100 and more genomes, that comes with an easy-to-use graphical user interface. The underlying gene cluster model can cope with deletions, insertions, as well as inversions. Additionally, Gecko 2 features a sound statistical evaluation.

Taxy-Pro: mixture modelling of metagenomes based on protein domain frequencies

Heiner Klingenberg, Kathrin Petra Aßhauer, Thomas Lingner and Peter Meinicke
Bioinformatics Department, University of Göttingen
peter@gobics.de

A central task in metagenomics is to estimate the phylogenetic composition of a sample from the extracted DNA. Recently, the Taxy mixture model [MAL11] has been introduced as a novel approach to taxonomic profiling of metagenomic sequence data. In contrast to all previous methods Taxy is not based on taxonomic classification of sequence fragments but instead approximates the overall oligonucleotide distribution of a sample by a mixture of reference distributions from known genomes. We here introduce Taxy-Pro, a variant of the mixture model that approximates the overall protein domain distribution instead of the oligonucleotide frequencies of the original Taxy approach. Restricting the analysis to valid Pfam domain hits as obtained from the CoMet web-server for functional profiling [LASM11], Taxy-Pro estimates are more robust with respect to a decreasing sequence quality and still faster than with any sequence classification method. In comparison with Taxy and other oligonucleotide-based methods, Taxy-Pro shows a better generalization performance in situations where the reference genomes in current databases can only yield a limited coverage of the metagenome. In particular, our results indicate that Taxy-Pro can be used to provide fast and reliable estimates of viral and archaeal fractions in metagenomic data.

References

[LASM11] Thomas Lingner, Kathrin Petra Aßhauer, Fabian Schreiber, and Peter Meinicke. CoMet–a web server for comparative functional profiling of metagenomes. *Nucleic Acids Research*, 39(Web Server issue):W518–523, July 2011.

[MAL11] Peter Meinicke, Kathrin Petra Aßhauer, and Thomas Lingner. Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics (Oxford, England)*, 27(12):1618–1624, June 2011.

The overall structure of the amyloid precursor protein

Ina Coburger, Sven O. Dahms and Manuel E. Than

Leibniz Institute for Age Research – Fritz Lipmann Institute (FLI), Protein Crystallography Group, Beutenbergstr. 11, 07745 Jena, Germany

ikoennig@fli-leibniz.de

The amyloid precursor protein (APP) is a type I transmembrane protein that is expressed in a wide number of different cell types. Proteolytic processing by beta- and gamma-secretases releases 38-43 amino acid long peptides, so called Aβ amyloid peptides that accumulate within the plaques in the brain of Alzheimer´s disease patients. Alternatively, initiation of the proteolysis cascade by alpha-secretase prevents the development of these toxic peptides.

In spite of intense research regarding the involvement of APP in Alzheimer´s disease, the three-dimensional structure of the entire protein, its physiological function and the regulation of its proteolytic processing remain largely unclear. To gain a deeper understanding about it, we cloned and recombinantly expressed different constructs of APP in *E. coli*. Using limited proteolysis and analytical gel permeation chromatography coupled static light scattering we experimentally determined that the large ectodomain of APP consists exactly of two rigidly folded domains – the E1- and the E2-domain. The acidic domain, connecting E1 and E2, as well as the juxtamembrane region, connecting E2 to the single transmembrane helix, are highly flexible and extended. Using analytical gelfiltration experiments and pull-down assays we further analysed whether the E1- and E2-domains interact with each other.

References

1. Matthias Gralle and Sergio T. Ferreira. Structure and functions of the human amyloid precursor protein: the whole is more than the sum of its parts. *Prog Neurobiol,* 82:11-32, 2007.
2. Sven O. Dahms, Sandra Hoefgen, Dirk Roeser, Bernhard Schlott, Karl-Heinz Guhrs, Manuel E. Than. Structure and biochemical analysis of the heparin-induced E1 dimer of the amyloid precursor protein. *Proc Natl Acad Sci U S A,* 107:5381-5386, 2010.

Heparin dependent dimerization of APP is mediated by its E1 but not its E2 domain

Sandra Hoefgen, Sven O. Dahms, Dirk Roeser and Manuel E. Than

Leibniz Institute for Age Research – Fritz Lipmann Institute (FLI) Jena, Protein Crystallography Group, Beutenbergstr.11, 07745 Jena, Germany

shoefgen@fli-leibniz.de

Alzheimer's disease is one of the most frequent forms of dementia in the elderly population affecting about 25 % of people in the age of 80 to 90 years. Due to the more and more ageing society the importance of dementia is increasing.

The brain of affected patients is characterized by the deposition of senile plaques containing the neurotoxic peptide $A\beta_{40-42}$ that is derived from its precursor, the Amyloid Precursor Protein (APP) [1]. Beside its role in Alzheimer's pathology many physiological functions are discussed for APP. However, until now it was not possible to correlate the known structures of subdomains with most of the proposed physiological functions of APP.

Recently, we could solve the structures of the APP-E1-domain, that dimerizes in a heparin dependent manner, [2] and of the metal binding APP-E2-domain [3]. Using wildtype proteins of both subdomains and a mutated E1-protein we could now show that the heparin induced dimerization of the APP-ectodomain is mediated by the E1- but not by the E2-domain.

References

[1] Dennis J Selkoe. Alzheimer's Disease: Genes, Proteins, and Therapy. *Physiological Reviews*, 81:741-766, 2001

[2] Sven O. Dahms, Sandra Hoefgen, Dirk Roeser, Bernhard Schlott, Karl-Heinz Gührs, Manuel E. Than. Structure and biochemical analysis of the heparin-induced E1 dimer of the amyloid precursor protein. *PNAS*, 107:5281-5387, 2010

[3] Sven O. Dahms, Ina Könnig, Dirk Roeser, Karl-Heinz Gührs, Magnus C. Mayer, Daniela Kaden, Gerd Multhaup, Manuel E. Than. Metal Binding Dictates Conformation and Function of the Amyloid Precursor Protein (APP) E2 Domain. *JMB*, 416(3):438-452, 2012

Visualization of the sensitivity of BLAST to changes in the parameter settings

Svenja Simon[1], Daniela Oelke[2], Klaus Neuhaus[3], Daniel A. Keim[1]
[1]*Data Analysis and Visualization Group, University of Konstanz*
[2]*German Institute for Educational Research and Educational Information, Frankfurt*
[3]*Microbial Ecology, Technical University Munich*
simon@dbvis.inf.uni-konstanz.de

BLAST [A$^+$90, A$^+$97] is a widely used algorithm in biology. Although, it is known that changes of the parameter settings result in differing result lists, in practice mostly default settings are used. These are appropriate in many cases but do not work well in all cases. In this work, we systematically study the sensitivity of BLAST results to changes in parameter settings for a broad range of input sequences. To compare the result lists, we use a pixel-based visualization technique. This technique permits a compact representation of changes in order and occurrence of subject sequences ("hits") in the result lists and also a high scalability to enable the comparison of many results at once. To include meta-information, e.g., the information if a hit is below a specific e-value threshold, we use visual boosting techniques [O$^+$11]. First results show that the sensitivity of BLAST depends largely on the input sequence and that the results differ a lot for changed parameter settings in some cases. To enable rapid insights, we will extend our approach towards a visual analytics system to analyze the relation between the parameter space and result lists of BLAST.

Keywords: BLAST, pixel-based visualization, visual analytics

References

[A$^+$90] Stephen F. Altschul et al. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[A$^+$97] Stephen F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[O$^+$11] Daniela Oelke et al. Visual Boosting in Pixel-based Visualizations. *Computer Graphics Forum*, 30(3):871–880, 2011.

Repeat identification and annotation in next generation sequencing data of complex eukaryotic genomes without a reference sequence.

Philipp Koch, Bryan Downie, Kathrin Reichwald and Matthias Platzer
Genome Analysis, Leibniz Institute for Age Research – Fritz Lipmann Institute, Jena, Germany
philippk@fli-leibniz.de

Many *de novo* sequencing efforts of complex eukaryotic genomes using next generation sequencing (NGS) are hampered by a large repeat content. For example nearly 50% of the human, 55% of the zebrafish and 80% of the barley genome are composed of repeats. Reads from current NGS technologies are too short to span most of the repeats. Resulting graph ambiguity in the assembly process is a major factor influencing both assembly completeness and coherency. To address these challenges, we set out to develop a method specifically treating the repetitive content of complex vertebrate genomes within NGS data sets. Based on k-mer counting, we first mask the unique segments within the original reads. After masking, repeat containing reads are *de novo* assembled (CLC assembly cell), producing a catalogue of repeat consensi. Subsequently, these families are being analyzed with respect to completeness and annotation. As proof of principle, we simulated paired-end reads of the *Drosophila melanogaster* genome (40x genome coverage). After masking and *de novo* assembly using different k-mer frequency thresholds, we analyzed each assembly for both completeness of *D. melanogaster* repeat families and for overall repeat content. The best repeat catalogue currently produced by this method contains 759 contigs (939 bp average contig size) with a repeat content of 85%, representing 196 out of 249 (79%) *D. melanogaster* repeat consensi annotated in RepBase Update (version 20120418). After further optimization and validation, this method will be applied to the genome of the short-lived killifish *Nothobranchius furzeri*, a new model organism for age research [Gen05]. This genome has an estimated size of 1.6-1.9 Gb [Rei09], for which initial analyses show a repeat content of 64% (21% of which are tandem repeats). Using the proposed strategy, we aim to construct and annotate a repeat library for this genome.

References

[Gen05] Genade T, et al. A: Annual fishes of the genus *Nothobranchius* as a model system for aging research. *Aging Cell 2005*, 4:223-233.

[Rei09] Reichwald K, et al. High tandem repeat content in the genome of the short-lived annual fish *Nothobranchius furzeri*: a new vertebrate model for aging research. *Genome Biol* 2009, 10:R16.

New network topology approaches reveal differential correlation patterns in breast cancer

Jan Budczies, Michael Bockmayr, Frederick Klauschen and Carsten Denkert
Institute of Pathologie, Charité – Universitätsmedizin Berlin
jan.budczies@charite.de

Analysis of genome-wide data is often carried out using standard methods such as differential expression analysis and clustering analysis. Beyond that, differential correlation analysis was suggested to identify changes in the correlation patterns between disease states [KS04, TBJ10]. The detection of differential correlation is a demanding, as the number of entries in the gene-by-gene correlation matrix is large. Currently, there is no gold standard for the detection of differential correlation. Therefore, we developed two untargeted algorithms (DCloc and DCglob) that identify differential correlation patterns by comparing the local or global topology of correlation networks. Transition from a correlation structure to a network requires fixing of a correlation threshold. Instead of a single cutoff, the new algorithms systematically investigate a series of correlation thresholds and permit to detect different kinds of correlation changes at the same level of significance: strong changes of a few genes and moderate changes of many genes. Lists of differentially correlated genes were compiled including false discovery rate (FDR) estimation by a random subsampling method. As use case, we compared the correlation structure of estrogen receptor negative (ER-) and estrogen receptor positive (ER+) breast cancer. Using DCloc, 770 differentially correlated genes were detected with a FDR of 12.8%. Using DCglob, 630 differentially correlated genes were detected with a FDR of 12.1%. Patterns of genes that were strongly correlated in ER- compared to ER+ breast cancer included clusters of marker genes for invasive apocrine breast cancer and for androgen receptor (AR) responsive breast cancer. The analysis of the estrogen receptor status in breast cancer illustrated that the network topology based approach for differential correlation can help to identify biologically relevant gene patterns beyond classical differential expression analysis.

References

[KS04] D Kostka and R Spang. Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20 (suppl 1):i194-i199, 2004.
[TBJ10] B Tesson, R Breitling and R Jansen. DiffCoEx: a simple and sensitive method to find differentially co-expressed gene modules. *BMC Bioinformatics*, 11:497, 2010.

A discriminative approach for finding motifs in ChIP-seq data

Jens Keilwagen[1], Ivo Grosse[2], Stefan Posch[2], and Jan Grau[2]
[1]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben*
[2]*Institute of Computer Science, Martin Luther University Halle–Wittenberg, Halle*
jens.keilwagen@ipk-gatersleben.de

Transcription factors are a main component of gene regulation as they bind to specific binding sites in promoters of genes and subsequently activate or repress gene expression. The de-novo discovery of transcription factor binding sites in target regions obtained by wet-lab experiments is still a challenging problem in bioinformatics, and has not been fully solved yet. Today, one major source of experimental data is chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-seq). Although this technique yields information about approximate binding regions, the exact motif positions and the motif itself still need to be determined computationally.

Here, we present a de-novo motif discovery approach specially tailored to ChIP-seq data. In contrast to previous approaches, it makes use of the binding profiles obtained from ChIP-seq and at the same time learns the binding motif discriminatively. While the binding profiles represent experimental binding probabilities that guide the approach towards regions with putatively high binding site abundance, the discriminative approach ensures that the discovered motif is specific to the ChIP-seq binding regions in contrast to genomic background. Due to the immense amount of sequence data obtained from ChIP-seq experiments, this novel approach is especially designed for an acceptable runtime on several 10,000 ChIP-seq positive regions. We demonstrate the utility of this approach on several publicly available ChIP-seq data sets for a variety of transcription factors.

Computational prediction of TAL effector target sites

Jan Grau[1], Annett Wolf[1], Stefan Posch[1], and Jens Boch[2]
[1]*Institute of Computer Science, Martin Luther University Halle–Wittenberg*
[2]*Dept. of Genetics, Institute of Biology, Martin Luther University Halle–Wittenberg*
jan.grau@informatik.uni-halle.de

Transcription activator-like (TAL) effectors are translocated into host plant cells by *Xanthomonas* bacteria via the type III secrection system, where they act as transcriptional activators. The DNA binding domain of TAL effectors is composed of conserved repeat structures with predictable binding specificity determined by repeat-variable diresidues (RVDs). Here, we present TALgetter, a new approach for predicting TAL effector target sites based on a statistical model, which represents *binding specificity* and *relevance* of RVDs independently. In contrast to previous approaches, the parameters of TALgetter are computationally estimated from training data.

We demonstrate that TALgetter is able to predict known TAL effector target sites and yields an improved prediction performance compared to the currently available approach. We study the binding specificities estimated by TALgetter and discover that different RVDs are differently relevant for the overall binding affinity. In subsequent studies, the predictions of TALgetter elicit a previously unreported positional preference of TAL effector target sites relative to the transcription start site. In addition, these studies reveal that the promoters of TAL effector target genes often contain a canonical TATA-box or TC-box, and that several TAL effector directly bind to the TATA-box, which might constitute one general mode of transcriptional activation by TAL effectors.

Scrutinizing the predictions of TALgetter, we identify novel putative TAL effector target sites in rice and sweet orange that are supported by microarray data and by functional analogy to known TAL effector target genes or a biological function related to pathogen infection. Hence, these TAL effector target sites are promising candidates for experimental validation. TALgetter is implemented as part of the open-source Java library Jstacs, and is freely available as a web-application and a command line program at `http://www.jstacs.de/index.php/TALgetter`.

Sequencing Copolymers using Mass Spectrometry

Martin S. Engler[1], Kerstin Scheubert[1], Sarah Crotty[2],
Ulrich S. Schubert[2,3] and Sebastian Böcker[1,3]
[1]*Chair of Bioinformatics,* [2]*Laboratory for Organic and Macromolecular Chemistry, Jena University Germany;* [3]*Jena Center for Soft Matter*
martin.engler@uni-jena.de

Synthetic polymers are an integral part of many modern materials with a vast application range in medicine, (bio-)chemistry, physics and material sciences. However, traditional characterization techniques do not provide information on the polymer chain, which is vital for understanding the relationship between the structure and the behavior of polymers - similar to *de novo* sequencing in proteomics.
So far mass spectrometry in polymer science is only used to gather information on end groups and molar masses, so the approach of sequencing polymers with MS and MS/MS spectra suggests itself. However, contrary to peptides the spectra of polydisperse polymers are very complex, because they contain mixtures of all possible molecules in different isomer classes. Up to date homopolymer MS/MS spectra can be analyzed either with rule-based[TJW+07] or *de novo*[BSP+11] algorithms. Copolymer spectra however are significantly more complex.
We present a work in progress of a new method which enables the analysis of synthetic copolymers independently of their polymer class or architecture. We propose a statistical model of the polymerization process and reduce the model fitting to several high-dimensional numerical optimization and NP-hard combinatorial problems and present preliminary results.

References

[BSP+11] A. Baumgaertel, K. Scheubert, B. Pietsch, K. Kempe, A. C. Crecelius, S. Böcker, and U. S. Schubert. Analysis of different synthetic homopolymers by the use of a new calculation software for tandem mass spectra. *Rapid Commun. Mass Spectrom.*, 25(12):1765–1778, 2011.

[TJW+07] K. Thalassinos, A. T. Jackson, J. P. Williams, G. R. Hilton, S. E. Slade, and J. H. Scrivens. Novel software for the assignment of peaks from tandem mass spectrometry spectra of synthetic polymers. *J. Am. Soc. Mass Spectrom.*, 18:1324–1331, 2007.

Seeking, sneaking and cheating: A game theoretical and spatially explicit modeling approach of life-history-strategies in Mucorales

Sarah Werner[1], Sebastian Germerodt[1], Patrick Faßbender[1], Anja Schroeter[1], Christine Schimeck[2], Johannes Wöstemeyer[2], Stefan Schuster[1]

[1] *Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany*

[2] *Institute of General Microbiology and Microbial Genetics, Friedrich Schiller University Jena, Neugasse 24, D-07743 Jena, Germany*

Sarah.Werner@uni-jena.de

Besides asexual reproduction, *Mucorales* show mating behavior and some species are also able to parasitize other mucoralean fungi. Mating as well as parasitism is triggered by the pheromone trisporic acid (TA) and its precursors. This pheromone is used to find a mating partner or host. By assuming that different strategies generate different net-payoffs we examine the role of resource availability as a potential trigger to resolve the question: when is which strategy - to mate or to parasitize - best?

We model the situation based on evolutionary game theory. Simulations show that mating is the preferable strategy for high resource availability or low parasitizing efficiency. Parasitism pays off if resources are scarce or if it is efficient enough. For very low resource availability it is the best not to interact at all.

Spatial and stochastic effects which are disregarded in the game-theoretic model are considered in an extended model - a cellular automaton. This automaton contains the rules of the game and additionally takes into account that resources are spatially heterogeneously distributed and are consumed by the fungi, represented by cells in a cartesian grid.

The automaton reveals a temporal sequence of the preferred strategies. This also leads to a spatially heterogenous distribution of the strategies reflecting the local availability of resources and the specific composition of the neighborhood.

Our results demonstrate that for *Mucorales*, which live in environments with permanently changing resource availability, it makes sense to have a broad repertoire of strategies, thus, being flexible and able to adjust to changing conditions.

Omix - A Tool for Customizable Visualization in the Context of Metabolic Networks

Peter Droste, Wolfgang Wiechert, Katharina Nöh
IBG-1: Biotechnologie & JARA - HPC, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
{p.droste,w.wiechert,k.noeh}@fz-juelich.de

Visualization tools for biochemical network diagrams guide our perception and understanding of wired cellular interactions. These tools usually link specific data from databases, experiments and simulations to the networks' elements. We present Omix, a software tool for flexible mapping and combination of any type of data in one single network diagram. The specially tailored scripting language Omix Visualization Language OVL gives a fast and intuitive access to all visual properties of nodes and edges composing the network by making them readily programmable by the user. Further significant features of Omix are:

- displaying network diagrams in different levels of detail depending on the study under focus.
- the "Visualization on Demand" (VoD) technique for user-controlled display of component-related information in different levels of detail. This technique avoids overloading of the network diagram while maintaining all available information.
- a semi-automatic layout approach to quickly draw large-scale metabolic networks according to historically evolved, well-established layout standards.
- extensibility by plug-ins. Plenty of plug-ins are available for compatibility, database access, network modelling and analysis, as well as the presentation of simulation results in 2D and 3D.

By the novel concept of flexible network programmability, Omix stands out of the family of existing visualization tools. Visualization in the context of metabolic networks becomes highly adaptable to different fields of application and individual requirements. The efficient and scalable visual exploration of

multi-omics data enables scientists to discover novel insights and derive hypotheses about cellular processes and biological mechanisms. Supplementary information and software download can be found on www.13cflux.net/omix.

References

Droste *et al*. LNCS vol. 6026, 163-174, 2010
Droste *et al*. Biosystems, 105 (2), 154-161, 2011
Droste *et al*. InfoVis 11 (3), 171-187, 2012

Predicting Ordinal Therapy Response with High-Dimensional Expression Data

Andreas Leha* and Klaus Jung and Tim Beissbarth
Department of Medical Statistics, University Medical Center Gttingen
andreas.leha@med.uni-goettingen.de

Molecular diagnosis or prediction of clinical treatment outcome based on high-throughput genomics data is a modern application of machine learning techniques for clinical problems. In practice, clinical parameters, such as patient health status or toxic reaction, are often measured on an ordinal scale (e.g. *good*, *fair*, *poor*).

Commonly, the prediction of ordinal end-points is treated as a multi-class classification problem, disregarding the ordering information contained in the response. This may result in a loss of prediction accuracy. Classical approaches to model ordinal response directly, including for instance the cumulative logit model, are typically not applicable to high-dimensional data.

Although there have been some extensions of existing methods for response prediction tailored towards ordinal response and high-dimensional data (e.g. [AW12]), the choice of methodology is still limited and the field is still lacking a comparative study.

We present a comparison of several approaches for ordinal classification on real world data as well as simulated data including the novel algorithm *hierarchical twoing (hi2)* that extends [FH01] and combines the power of well-understood binary classification with ordinal response prediction.

Our findings suggest, that the classification performance of an algorithm is dominated by its ability to deal with the high-dimensionality of the data. Although the comparative evaluation do not show a clear winner, taking the ordinality of the response into account can improve the classification accuracy.

References

[AW12] K.J. Archer and A.A.A. Williams. L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine*, 2012.

[FH01] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *In: Proc 12th Europ Conf on Machine Learning*, pages 145–156. Springer, 2001.

Identification of highly diverse genomic regions in German Holstein dairy cattle

R. H. Bortfeldt, A. O. Schmitt, G. A. Brockmann
Breeding Biology and Molecular Genetics, Humboldt University of Berlin
ralf.bortfeldt@agrar.hu-berlin.de

Holstein dairy cattle have been selected for more than a century with the aim to increase total milk yield but also milk constituents such as milk fat and milk protein. Besides the ability for high milk production it is important to improve the fitness of high performing cows. It is an established hypothesis that fitness related genes may be particularly located in genomic regions of enhanced variability [1]. Based on haplotype block structures, which combine single nucleotide polymorphisms (SNPs) that are inherited more often together, frequencies of heterozygous SNPs and different haplotypes provide measures to assess the population variability in the respective genomic region. Together with known fitness traits, statistical association can be used to identify regions with significant effects on these traits.

We used a genome wide set of 43,686 SNPs typed with the Illumina BovineSNP50K Beadchip [2] in approximately 2,400 German Holstein breeding bulls to establish a haplotype block structure, filter regions with excess of variability and associate them with breeding values for life expectancy, fertility and somatic cell content of the milk as representative fitness traits.

We show that several bovine chromosomes maintain regions with above-average diversity by means of heterozygosity and haplotype inventory. These regions are more frequently significantly ($p < 0.05$) associated with the analysed fitness traits than with production traits that are currently in the focus of selection. We suggest allelic variants of candidate genes located in these haplotype blocks as putative markers for further investigation and inclusion in sustainable selection schemes.

References

1 Traherne, J.A., *et al.* (2006) Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet* 2, e9

2 Matukumalli, L.K., *et al.* (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* 4, e5350

Rhythm of epigenetics: dancing to the bead of DNA methylation

Flemming S.[1], Grüning B.[1] , Bohleber S.[1] , Häupl T.[2] , Günther S. [1]
stephan.flemming@pharmazie.uni-freiburg.de

[1] *Institute of Pharmaceutical Sciences, University of Freiburg*
[2] *Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin*

Methylation of cytosins within a CpG dinucleotide is a common epigenetic DNA modification and may lock cells in a pathogenic state in complex disorders, like cancer [1] or rheumatoid arthritis [2]. CpGs occur mainly in clusters, so called CpG islands (CPIs) and nearly 70% of the human genes have CPIs in their promotor region [3].

The Illumina HumanMethylation450 Beadchip platform provides a genome-wide coverage of 480 000 CpGs. Analysis of these CpGs reveals a correlation between changes in DNA methylation and gene expression but it seems that not all sites have the same impact.

The metyhlation state of one CpG or a whole CPI may influence the expression of the corresponding gene due to binding of Methylation-Binding-Domains (MBD) and other methylation depended proteins.

To identify CpGs influencing gene expression and to identify methylation patterns we used several approaches, e.g. network analysis and machine learning techniques. The results help to understand differential methylation analyses with Illumina BeadChip platform and other techniques.

References

[1] Hansen KD *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet.*, 43(8):768, 2011
[2] Karouzakis E et al. DNA methylation regulates the expression of CXCL12 in rheumatoid arthritis synovial fibroblasts. *Genes and Immunity,* 12:643–652, 2011
[3] Nakano K *et al.* DNA methylome signature in rheumatoid arthritis. *Ann Rheum Dis*., ahead of print, 2012

Improving Fragmentation Tree Alignments by Joining Fragmentation Events

Kai Dührkop, Sebastian Böcker
Chair of Bioinformatics, Friedrich-Schiller-University Jena
kai.duehrkop@uni-jena.de

Mass spectrometry is a key technology for sensitive, automated and high-throughput analysis of small molecules such as metabolites. Recently, fragmentation trees have been introduced for the automated analysis of their fragmentation patterns [BR08]. Similarities of fragmentation trees using unordered tree alignments correlate with chemical similarities of their compounds [RSH$^+$12]. This technique allows to identify small molecules even if they are not contained in any database. Although the unordered tree alignment problem is computationally hard, an exact dynamic programming algorithm has been introduced which solves the problem swift in practice [HDR$^+$12].

Fragmentation trees computed from spectra measured on different instruments or with different parameter settings may differ in the number of detected fragmentation events. In the resulting trees, a single loss in one tree can be a combination of several losses in the other tree. The classical tree alignment approach has difficulties to compare these trees. We extend the unordered tree alignment problem with a multi-join operation which aligns multiple fragmentation events at once. It can handle missing fragmentation events as well as different order of fragmentation steps. Extending the dynamic programming algorithm with this new operation increases the runtime quadratically. We evaluate the benefits of this method and present preliminary results.

References

[BR08] Sebastian Böcker and Florian Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:I49–I55, 2008. Proc. of *European Conference on Computational Biology* (ECCB 2008).

[HDR$^+$12] Franziska Hufsky, Kai Dührkop, Florian Rasche, Markus Chimani, and Sebastian Böcker. Fast alignment of fragmentation trees. *Bioin-*

formatics, 28:i265–i273, 2012. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2012).

[RSH+12] Florian Rasche, Kerstin Scheubert, Franziska Hufsky, Thomas Zichner, Marco Kai, Aleš Svatoš, and Sebastian Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal Chem*, 84(7):3417–3426, 2012.

Increasing the quality of FlipCut supertrees

Markus Fleischauer and Sebastian Böcker
Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena
markus.fleischauer@uni-jena.de

Supertree methods as part of a divide-and-conquer approach can help to enhance the reconstruction of large phylogenies in speed and quality. We compute maximum likelihood (ML) trees for several subsets of taxa. Then we merge the resulting trees into a supertree which is used as seed for a ML analysis on the complete dataset. Current sufficiently accurate supertree methods are too slow for such an approach. Brinkmeyer *et al.* [BGB11] developed a polynomial supertree method called FlipCut supertrees, which computes supertrees of high quality. We will further advance FlipCut so that the quality of the supertrees is on par with Matrix Representation with Parsimony (MRP). We pursue two different approaches to increase the quality of FlipCut supertrees. For one thing we improve the scoring of the FlipCut graph because it has much influence of the supertree quality. We use distances (path lengths) between the taxa of the source trees, to resolve ambiguous information of the tree topology. For another thing we developed a beam search algorithm, to consider suboptimal cuts across the FlipCut graph. This minimizes the sum of costs of all minimum cuts executed during the FlipCut algorithm.
We found out, that the beam search algorithm admittedly reduces the cut-costs, but not appreciably improves the supertree quality. However, the new type of cost function even in its initial version increases the supertree quality.

References

[BGB11] Malte Brinkmeyer, Thasso Griebel, and Sebastian Böcker. FlipCut Supertrees: Towards Matrix Representation Accuracy in Polynomial Time. In *Proc. of Computing and Combinatorics Conference (COCOON 2011)*, volume 6842, pages 37–48, 2011.

Using Metabolic Modelling and Optimization Methods in Organ-oriented Systems Biology: Prediction of Adaptive Liver Zonation during Regeneration

Martin Bartl[1,2], Michael Pfaff[2,3], Dominik Driesch[2], Sebastian Zellmer[4,5], Stefan Schuster[6], Rolf Gebhardt[4], Pu Li[1]

1 *Institute for Automation and Systems Engineering, Ilmenau University of Technology,* {martin.bartl; pu.li}@tu-ilmenau.de

2 *BioControl Jena GmbH,* dominik.driesch@biocontrol-jena.com

3 *Department of Medical Engineering and Biotechnology, University of Applied Sciences Jena,* michael.pfaff@fh-jena.de

4 *Institute of Biochemistry, University of Leipzig,* rolf.gebhardt@medizin.uni-leipzig.de

5 *Present Affiliation: Department: Safety of Consumer Products, Federal Institute for Risk Assessment, Berlin,* sebastian.zellmer@bfr.bund.de

6 *Department of Bioinformatics, Friedrich Schiller University Jena,* stefan.schu@uni-jena.de

As a highly structured organ, the liver is consisting of a large number of identical subunits, the lobuli. The metabolism within the individual lobules is known to be zonated, i.e. certain metabolic reactions take place at specific locations in the lobule. This applies e.g. to nitrogen, carbohydrate and lipid metabolism. In order to support organ-oriented, spatio-temporal modelling of the liver, essential reactions of the zonated nitrogen metabolism were described and predicted using a combined metabolic modelling and optimization approach. It was originally applied to the healthy liver and based on this - as reported here - to the regenerating liver after CCl_4 intoxication. This study yielded the prediction of a phenomenon that could be called 'adaptive' or 'dynamic' zonation - i.e. changing patterns of enzyme distributions along the liver lobule acinus during regeneration until the healthy state is restored.

More specifically, the approach is based on an optimality criterion that represents liver functionality with respect to ammonia detoxification by ureogenesis and glutamine regulation, constrained by maximum enzyme capacities. The prediction indicates in the case of maximum damage that there is only high activity of carbamoyl phosphate synthetase, the key enzyme for ureogenesis. During the regeneration process, the enzymatic

activity of carbamoyl phosphate synthetase extends continuously with only slow development of glutaminase in the periportal zone. Only in the later stages of regeneration, glutamine synthetase is fully recovered in the pericentral zone.
In summary, a novel approach was established to investigate physiological strategies of the liver after intoxication resulting in plausible predictions of the regeneration process. The combination of metabolic models and optimization techniques provides a promising approach to identify unknown structures of liver zonation. The application of this approach to a higher detail of liver lobule representation, other zonated metabolic reactions or other organs has the potential to unravel further unknown phenomena.

Trancriptomic analysis of the polyplod adriatic sturgeon, *Acipenser nacarii*

Michele Vidotto[1], Alessandro Coppe[1], Abhishek Kumar[1,2], Alessandro Grapputo[1], Gilberto Grandi[1] and Leonardo Congiu[1]

[1]Biology Department, University of Padova, Padova, Italy
[2]Abteilung für Botanische Genetik und Molekularbiologie Botanisches, Institut und Botanischer Garten, Christian-Albrechts-Universität, Kiel, Germany

michele.vidotto@studenti.unipd.it |akumar@bot.uni-kiel.de

The present project exploits advance in next generation technology, to bridge the gap in the genomic resources of a very ancient and critically endangered with high evolutionary and economical interest, Sturgeons. Here we provide the first transcriptome characterization of *A. naccarii,* tetraploid species with an estimated number of 240 chromosomes. The potential of the 454 FLX sequencing technology, along with the titanium chemistry, was exploited for sequencing 2 normalized libraries from gonads and brain of 2 individuals: one male (cDNA3) and one female (cDNA4). After a preprocessing step reads from both libraries were tagged accordingly to the sex of origin and jointly assembled using MIRA assembler. After two iterative assembly runs a total of 55,282 sequences, 42,193 contigs (21.87 Mb) and 13,089 singletons (3.91 Mb) were realized. To evaluate the coverage of the cDNA libraries by reads, we performed a rarefaction analysis using zebrafish complete cDNA set as a reference. The extrapolation from the hyperbolic model, showed that we identified ~88% of the total different transcripts potentially identifiable at infinite sequencing. To estimate the total number of *A. nacarii* transcripts potentially present in the two tissues (gonads and brain), we adapted the capture-recapture method widely used in ecology to estimate animal population sizes, considering the two sequencing experiments as two independent samples from the same transcripts population. We estimated the transcripts population size to be 68.904, through the Rcapture. We assigned 8,784 contigs (16%) with GO term via Blast2GO annotation. TBLASTN comparison of contigs against the complete cDNA sets from other fish genomes showed that among teleosts, the fraction of putatively orthologous transcripts and remains more or less the same in the Acanthopterygians. FREEBAYES with probability cut-off 0.9 and 5 as minimum coverage threshold, identified 21,791 putative SNPs (94.04%) and 57,996 INDELs (98.05%) from 6,283 and 8,678 contigs respectively. Estimation of the non-synonymous (Ka) to synonymous (Ks) mutation rate, in contig containing

SNPs, revealed 0.03% of them under possible diversifying selection.

PIPS: Software to predict Pathogenicity Islands and analysis of genome plasticity in *Corynebacterium pseudotuberculosis*

Vinicius A. C. de Abreu[1*§], Siomar C. Soares[1*§], Vasco Azevedo[1] Jan Baumbach[2]

[1] General Biology Department, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

[2] Computational Systems Biology research group, Center for Bioinformatics, Saarland University - Germany

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

Vinicius de Abreu: vini.abreu@gmiail.com

Siomar Soares: siomars@gmail.com

Abstract

We have developed the software PIPS (Pathogenicity Island Prediction Software), aiming to predict PAIs using a novel and more complete approach based on the detection of multiple PAIs features. In contrast to other existing tools, our approach is capable of utilizing multiple features for pathogenicity island detection in an integrative manner. Additionally, besides the automatically generated list of putative PAIs, PIPS also generates files to perform manual curation of the data. Finally, we validated PIPS and show it provides better accuracy rates than the other available software packages results After validation, we used PIPS to perform an analysis on the genome sequence of *Corynebacterium pseudotuberculosis* and the software identified 14 putative PAIs (automatic analysis). After manual curation, a comparative analysis of 7 PAIs was realized to refine the results. The 7 PAIs harbour genes such as fag A, B, C, D operon (iron acquisition) and pld gene (*sphingomyelinase action*). Additional analysis were made through comparison of *C. Pseudotuberculosis* and *C. diphtheriae* . As a result, we found PiCp3 of *C. Pseudotuberculosis* is also present in *C. Diphtheriae* (PiCd3) and PiCP6 of *C. Pseudotuberculosis* is substituted by another PAI in *C. Diphtheriae* (PICD8). PIPS is an open source and easy to install software, which presents a high efficiency in the identification of putative PAIs *in silico*. The web based interface and the source code for PIPS, can be found on http://www.genoma.ufpa.br/lgcm/pips.

CRACPipe: UV crosslinking and analysis of cDNA pipeline

Stefan Simm, Roman Martin, Maike Ruprecht, Jens Einloft, Markus T. Bohnsack, Oliver Mirus,
Enrico Schleiff

Institute of Molecular Biosciences, JWGU Frankfurt am Main

stefan_simm2000@yahoo.de

Next Generation Sequencing (NGS) is a recent method for producing a high amount of sequence information. We have developed the software pipeline CRACPipe to analyze data derived from the NGS method called Solexa (Illumina).

The data were generated from different approaches of unbiased UV-crosslinking and analysis of cDNA (CRAC) experiments [BM09]. With this experimental method protein-RNA interaction sites can be identified and mapped on the complete genome of an organism. *In vivo* or *in vitro* UV-crosslinking is followed by purification of crosslinks via a tag on the protein of interest and negative controls are performed by UV-crosslinking cells that do not express a tagged protein. After digestion of the protein, ligation of linkers, reverse transcription and PCR, the samples are analysed by Solexa NGS. The millions of short reads are preprocessed and then mapped onto the complete genome via SSAHA2 [NC01]. The mapped reads can be allocated to different feature types on the genome, as e.g. genes or rDNA.

In the end our method creates histograms and calls peaks for visualization via MochiView [HJ10]. The histograms show the number of mapped reads for each position on the chromosome and the mutations in the mapped reads. Mutations allow the identification of UV-crosslinking sites. In addition to these histograms an overview of aligned reads and the occurrence of mapped reads in the different feature-types is generated. Non-annotated regions of the genome or areas between annotated features are also included in the alignment and possibly represent new putative binding sites.

For the ribosome of yeast and the RDN37-1 and RDN5-1, these transcripts were aligned against and modeled into the crystal structure of bacterial ribosome. With this sequence-to-structure alignment it is possible to map the afore mentioned histograms onto the structure. By visual inspection of the ribosome structure colored according to the occurrence of mapped reads and mutations, we can localize putative binding sites in 3D.

References

[BM09] Markus T. Bohnsack, Roman Martin, Sander Granneman, Maike Ruprecht, Enrico Schleiff, David Tollervey. Prp43 Bound at Different Sites on the Pre-rRNA Performs Distinct Functions in Ribosome Synthesis. *Molecular Cell*, 36(4): 583-592, 2009.

[NC01] Zemin Ning, Anthony J. Cox, James C.Mullikin. SSAHA: a fast search method for large DNA databases. *Genome Res*., 11(10): 1725–1729, 2001.

[HJ10] Oliver R. Homann and Alexander D. Johnson. MochiView: versatile software for genome browsing and DNA motif analysis. *BMC Biology*, 8:49, 2010.

Tom40 - An Outer Membrane Protein

Nadine Flinner, Enrico Schleiff and Oliver Mirus
Institute of Molecular Biosciences, JWGU Frankfurt am Main
nadine-flinner@gmx.de

Tom40 is a β-barrel Protein in the outer membrane of mitochondria and forms the central translocation pore of the TOM (Translocase of the Outer membrane of Mitochondria) complex. Most of the mitochondrial proteins are encoded in the nucleus and are translated in the cytosol. The TOM complex is required for the recognition and import of these proteins and consists of several additional subunits: the three small Tom proteins with regulatory functions: Tom5, Tom6 and Tom7, the receptor component Tom22 and two additional peripheral receptors Tom20 and Tom70. In general, Tom22 and Tom20 recognize proteins with an amphipathic presequence and Tom70 is responsible for those proteins, which do not carry a classical presequence and are bound to HSP proteins. [EY10]

Tom40 is homologous to VDAC (Voltage-Dependent Anion Channel) [PC09], which is a β-barrel with 19 β-strands and an N-terminal helix and is also located in the outer membrane of mitochondria [UC08]. Both proteins share little sequence similarity (Psi-Blast: *nc*Tom40 → 3rd round → *nc*VDAC (similarity: 33/180 aligned positions; E-Value 6e-5); *mm*VDAC (similarity: 50/305 aligned positions; E-Value 4e-29)), because on the one hand the residues exposed to the membrane are subject to a low selective pressure and on the other hand both proteins perform different functions.

We now aimed at building a reliable homology model. To do so we developed a new alignment strategy to get a reliable alignment between the sequence of the template structure (*mm*VDAC - pdb ID: 3emn) and the target sequence (*nc*Tom40). Subsequently, we used our homology model to identify conserved structural motifs in theTom40 family, which are distinct from the VDAC family and could therefore be important for the function of the protein.

References

[EY10] Toshiya Endo and Koji Yamano. Transport of proteins across or into the mitochondrial outer membrane. *Biochim Biophys Acta*, 1803(6):706-14, 2010.

[PC09] Mascha Pusnik, Fabien Charrière, Pascal Mäser, Ross F Waller, Michael J. Dagley, Trevor Lithgow, André Schneider. The single mitochondrial porin of Trypanosoma brucei is the main metabolite transporter in the outer mitochondrial membrane. *Mol Biol Evol*, 26(3):671-80, 2009

[UCA08] 1: Rachna Ujwal, Duilio Cascio, Jacques P. Colletier, Salem Faham, Jun Zhang, Ligia Toro, Peipei Ping, Jeff Abramson. The crystal structure of mouse VDAC1 at 2.3 A resolution reveals mechanistic insights into metabolite gating. *Proc Natl Acad Sci U S A*, 105(46):17742-7, 2008.