# GCB 2013 Göttingen - Poster Abstracts

## Contents

**Table of Contents** 1

**P1 SubtiWiki - more than just a Wiki**
*Raphael Michna and Jörg Stülke* 8

**P2 Wondolin: A Web Service for Protein Sequence Comparison**
*André Ahrens and Martin Kollmar* 9

**P3 WaggaWagga: a Web Service to Compare and Visualize Coiled-Coil Predictions, and to Assess the Oligomerisation State of Coiled-Coils**
*Dominic Simm, Klas Hatje and Martin Kollmar* 10

**P4 OPTIMAS-DW: Integrating Different Maize -Omics Information into a Data Warehouse**
*Christian Colmsee, Tobias Czauderna, Anja Hartmann, Martin Mascher, Jinbo Chen, Matthias Lange, Falk Schreiber and Uwe Scholz* 11

**P5 Definitions and nomenclatures for alternative splicing events**
*Martin Pohl and Stefan Schuster* 13

**P6 EndoNet: An information resource about the intercellular signaling network**
*Jürgen Dönitz and Edgar Wingender* 14

**P7 BRENDA in 2013: integrated strains, kinetic data, improved disease classification**
*Sandra Placzek and Dietmar Schomburg* 16

**P8 Combining ontology design and flexible data management in biomedical sciences.**
*Timm Fitschen, Alexander Schlemmer, Daniel Hornung, Philip Bittihn, Johannes Schröder-Schetelig, Ulrich Parlitz and Stefan Luther* 18

**P9 Comparison of platforms to integrate transcriptome data from different sources**
*Sarah N. Mapelli, Bjoern Schumacher, Ankit Arora and Karsten R. Heidtke* 19

**P10 Resource-Constrained Analysis of Ion Mobility Spectra with the Raspberry Pi**
*Elias Kuthe, Alexey Egorov, Alexander König, Marcel Köppen, Henning Kühn, Suzana Mitkovska, Marianna D'Addario, Dominik Kopczynski and Sven Rahmann* 21

**P11 rBiopaxParser: A new package to parse, modify and merge BioPAX-Ontologies within R**
*Frank Kramer, Michaela Bayerlova, Annalen Bleckmann and Tim Beissbarth* 22

**P12 MarVis-Graph for integrative analysis of metabolic reaction chains in non–targeted experiments**
*Manuel Landesfeind, Alexander Kaever, Kirstin Feussner, Ivo Feussner and Peter Meinicke* 23

**P13 The MarVis-Suite: Integrative and explorative analysis of Metabolomics and Transcriptomics data**
*Alexander Kaever, Manuel Landesfeind, Kirstin Feussner, Ivo Feussner and Peter Meinicke* **25**

**P14 FractalQC: A Bioconductor Package for Quality Control of RNA-Seq Coverage Patterns by Means of the Fractal Dimension**
*Stefanie Tauber and Arndt von Haeseler* **26**

**P15 Quantum Coupled Mutation Finder: Predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming**
*Mehmet Gültas, Martin Haubrock, Edgar Wingender and Stephan Waack* **27**

**P16 EnzymeDetector: Enhancements in 2013 for high throughput applications**
*Marcus Ulbrich and Dietmar Schomburg* **29**

**P17 BAM A tool for basic analysis of microarray data**
*Marc Bonin, Florian Heyl, Irene Ziska, Lydia Ickler, Denise Sinning, Jekaterina Kokatjuhha and Thomas Häupl* **31**

**P18 DeNovoGUI: an open-source graphical user interface for de novo sequencing of tandem mass spectra**
*Thilo Muth, Lisa Weilnböck, Erdmann Rapp, Christian Huber, Lennart Martens, Marc Vaudel and Harald Barsnes* **32**

**P19 mascR: Efficient NGS fragment-size estimation**
*Orr Shomroni and Stefan Bonn* **33**

**P20 New Network Analysis Tools Beyond Hairballs**
*Tim Kacprowski, Nadezhda T. Doncheva and Mario Albrecht* **35**

**P21 Identification of allele-specific expression and RNA editing sites in paired NGS data by ACCUSA2**
*Michael Piechotta and Christoph Dieterich* **36**

**P22 Dynamics of Two-Photon Two-Color Transitions in Flurophores Excited by Femtosecond Laser Pulses**
*Oleg Vasyutinskii, Karl-Heinz Gericke, Peter Shternin, Andrey Smolin, Sebastian Herbrich and Stefan Denicke* **38**

**P23 BiSQuID: Bisulfite Sequencing Quantification and Identification**
*Cassandra Falckenhayn, Guenter Raddatz and Frank Lyko* **39**

**P24 Detection and monitoring of excited biomolecules by means of holographic technique**
*Irina Semenova, Oleg Vasyutinskii and Alexandra Moskovtseva* **40**

**P25 A spherical model of alveolar macrophages using computerized graphical techniques**
*Dominic Swarat, Martin Wiemann and Hans-Gerd Lipinski* **41**

**P26 Computing metabolic costs of amino acid and protein production in Escherichia coli**
*Christoph Kaleta, Sascha Schäuble, Ursula Rinas and Stefan Schuster* **43**

**P27 The HGT Calculator: targeted detection of horizontal gene transfer from prokaryotes to protozoa in small data sets**
*Sabrina Ellenberger, Stefan Schuster and Johannes Wöstemeyer* **44**

**P28 Linking Phenotypes and Genomic Regions: the Forward Genomics Approach**
*Xavier Prudent and Michael Hiller* **45**

**P29 Automated combined analysis of DNA methylation and transcription profiles in different immune cells**
*Marc Bonin, Lorette Weidel, Stephan Flemming, Andreas Grützkau, Biljana Smiljanovic, Till Sörensen, Stephan Günther and Thomas Häupl* **46**

**P30 Automated Classification of Cell Populations with Multi-channel Flow Cytometry Data - Using Sparse Grids Classifying A Sparsely Populated Data Space**
*Manuel Nietert, Steve Wagner, Annalen Bleckmann, Klaus Jung, Dorit Arlt and Tim Beissbarth* **47**

**P31 Spatial distribution of cells in Hodgkin Lymphoma**
*Hendrik Schäfer, Tim Schäfer, Joerg Ackermann, Norbert Dichter, Claudia Döring, Sylvia Hartmann, Martin-Leo Hansmann and Ina Koch* **48**

**P32 Sequence based analysis of plant myosins**
*Stefanie Mühlhausen and Martin Kollmar* **49**

**P33 Transcriptome analysis of the model organism Tribolium castaneum**
*Sarah Behrens, Robert Peuß, Barbara Milutionivic, Hendrik Eggert, Daniela Esser, Philip Rosenstiel, Erich Bornberg-Bauer and Joachim Kurtz* **50**

**P34 Analysis of Wt1 ChIP Seq data from mouse glomeruli**
*Stefan Pietsch, Christoph Englert and Lihua Dong* **51**

**P35 Newtonian dynamics in the space of phylogenetic trees**
*Björn Hansen and Andrew E. Torda* **52**

**P36 Design of new inhibitors for HIV-Integrase: Implications of structure based drug design by Molecular Modeling approach**
*Jitendra Kumar Gupta, Nandhini K P, Annie Cynthia B, T. Gopala Krishnan and Asif Naqvi* **53**

**P37 Elucidating soil microbial communities in agricultural soils**
*Yudai Suzuki, Kazunari Yokoyama, Naomi Sakuramoto and Y-H Taguchi* **54**

**P38 A Novel Approach for Determining Spatial Colocalization of Proteins Inside Ceramide-rich Domains**
*Christian Imhäuser, Heike Gulbins, Erich Gulbins and Hans-Gerd Lipinski* **55**

**P39 Structure modeling of proteins for the biosynthesis of sex pheromones in zygomycetous fungi**
*Sabrina Ellenberger and Johannes Wöstemeyer* **57**

**P40 Functional and metabolic characterization of plant peroxisomal proteomes**
*Ana Tzvetkova, Sigrun Reumann, Peter Meinicke and Thomas Lingner* **59**

**P41 Different expression of classical Hodgkin lymphoma and primary mediastinal B-cell lymphoma**
*Denis Dalic, Ina Koch, Martin-Leo Hansmann and Claudia Döring* **60**

**P42 Statistical analysis of Hodgkin lymphoma based on tissue image data**
*Jennifer Scheidel, Tim Schäfer, Hendrik Schäfer, Jörg Ackermann, Claudia Döring, Sylvia Hartmann, Martin-Leo Hansmann and Ina Koch* **61**

**P43 Analysis of RNA-seq data for identifying flowering time regulators in vernalized and non-vernalized rapeseed**

*Claus Weinholdt, Nazgol Emrani, Ioana Lemnian, Nicole Jedrusik, Carlos Molina, Christian Jung and Ivo Grosse* **62**

**P44 Towards an optimal transcriptome assembly of the Naked Mole Rat**
*Martin Bens, Karol Szafranski and Matthias Platzer* **63**

**P45 Distinction of Type 2 diabetes using PCA, miRNA as features**
*Shodai Katsukawa and Y-H. Taguchi* **65**

**P46 Multiple Protein Alignment using Domain Information**
*Layal Al Ait and Burkhard Morgenstern* **66**

**P47 Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as drug target candidates**
*Ryoichi Kinoshita, Mitsuo Iwadate, Hideaki Umeyama and Y-H. Taguchi* **67**

**P48 A longitudinal transcriptome analysis of a fungal aging model indicates that autophagy compensates age-dependent proteasomal impairments**
*Oliver Philipp, Andrea Hamann, Jörg Servos, Alexandra Werner, Heinz D. Osiewacz and Ina Koch* **69**

**P49 Prediction of Methotrexate Treatment Response in Rheumatoid Arthritis via Affymetrix miRNA Microarray Profiling**
*Marc Bonin, Stephan Peter, Karsten Mans, Carolin Sohnrey, Gerd-Rüdiger Burmester, Thomas Häupl and Bruno Stuhlmüller* **71**

**P50 Chronic Inflammation is associated with cancer-related methylation changes**
*Sebastian Bender, Monther Abu-Remaileh, Günter Raddatz, Jehudit Bergman and Frank Lyko* **72**

**P51 Identify cell line specific microRNA TSS based on H3K4m3 data**
*Xu Hua, Jie Li, Jin Wang and Edgar Wingender* **73**

**P52 Protein Folding and Structure through Synchronization**
*Leandro Nadaletti, Beatriz Lima and Solange Guimarães* **74**

**P53 Network of Silence**
*Stephan Flemming, Simon Bohleber, Thomas Häupl and Stefan Günther* **75**

**P54 Prediciting Alzheimer Disease using miRNA signatures**
*Jerzy Dyczkowski, Pooja Rao, Angela Dettmar, Anja Schneider, Andre Fischer and Stefan Bonn* **76**

**P55 Genotype-phenotype correlation of continuous characters while considering phylogeny**
*Amol Kolte and Farhat Habib* **77**

**P56 Finding Functional Interactions of Proteins and Small Molecules in Sentences of PubMed Abstracts**
*Kersten Döring, Michael Becer and Stefan Günther* **79**

**P57 A parametric analyse of the asymmetric Wagner parsimony**
*Gilles Didier* **80**

**P58 Comparison of protein topology graphs using graphlet-based methods**
*Tatiana Bakirova, Tim Schäfer and Ina Koch* **81**

**P59 Predicting targets of synergistic microRNA regulation**
*Ulf Schmitz, Shailendra Gupta, Xin Lai, Julio Vera and Olaf Wolkenhauer* **82**

**P60 GEDEVO: An Evolutionary Graph Edit Distance Algorithm for Biological Network Alignment**
*Rashid Ibragimov, Maximilian Malek, Jiong Guo and Jan Baumbach* **84**

**P61 Nonlinear Methods of DNA Coding Regions Identification**
*Vyacheslav Tykhonov and Nataliia Kudriavtseva* **86**

**P62 Basic topological features for metabolic pathway models**
*Jens Einloft, Joerg Ackermann and Ina Koch* **87**

**P63 Prediction of protein interaction types based on sequence and network features**
*Florian Goebels and Dmitrij Frishman* **89**

**P64 Simultaneous Gene Prediction in Related Species**
*Stefanie König, Lizzy Gerischer and Mario Stanke* **90**

**P65 Multiple Protein Alignments – Structure Versus Sequence-Based**
*Iryna Bondarenko and Andrew E. Torda* **91**

**P66 RNA sequence design and experimental verification**
*Marco Matthies, Kristina Gorkotte-Szameit, Stefan Bienert, Cindy Meyer, Ulrich Hahn and Andrew Torda* **92**

**P67 Automated Peak Extraction for MCC/IMS Measurements of Exhaled Breath**
*Marianna D'Addario, Dominik Kopczynski, Jörg Ingo Baumbach and Sven Rahmann* **93**

**P68 Dinucleotide distance histograms for fast detection of rRNA in metatranscriptomic sequences**
*Heiner Klingenberg, Robin Martinjak, Frank Oliver Glöckner, Rolf Daniel, Thomas Lingner and Peter Meinicke* **94**

**P69 A Memory Efficient Data Structure for Pattern Matching in DNA with Backward Search**
*Dominik Kopczynski and Sven Rahmann* **95**

**P70 Mixture models for the estimation of metagenomic abundances**
*Kathrin P. Aßhauer, Heiner Klingenberg, Thomas Lingner and Peter Meinicke* **96**

**P71 Modelling NF-kB signal transduction using Petri nets**
*Leonie Amstein, Nadine Schöne, Simone Fulda and Ina Koch* **97**

**P72 Comparison of different graph-based pathway analysis methods on breast cancer expression data**
*Michaela Bayerlova, Frank Kramer, Klaus Jung, Florian Klemm, Annalen Bleckmann and Tim Beissbarth* **99**

**P73 NOVA: Evaluation of complexome profiling data**
*Heiko Giese, Jörg Ackermann, Heinrich Heide, Ilka Wittig, Ulrich Brandt and Ina Koch* **100**

**P74 Local Search for Bicriteria Multiple Sequence Alignment**
*Maryam Abbasi, Luis Paquete, Francisco Pereira and Sebastian Schenker* **102**

**P75 Boolean network reconstruction to explain individual drug response in breast cancer**
*Silvia von der Heyde, Christian Bender, Frauke Henjes, Johanna Sonntag, Ulrike Korf and Tim Beißbarth* **104**

**P76 Modeling and Simulation of Biological Networks using extended hybrid functional Petri nets**
*Christoph Brinkrolf, Sebastian Jan Janowski, Lennart Ochel, Martin Lewinski, Benjamin Kormeier, Bernhard Bachmann and Ralf Hofestädt* **106**

**P77 Detection of synergistic effects evoking new functions in a cell unsing a bipartite network algorithm**
*Sebastian Zeidler, Björn Goemann and Edgar Wingender* **108**

**P78 The $k$-Mismatch Average Common Substring approach**
*Chris Leimeister and Burkhard Morgenstern* **109**

**P79 High Betweenness – Low Connectivity (HBLC) Signatures in the Human Proteome**
*Thomas Wiebringhaus and Heinrich Brinck* **110**

**P80 An efficient approach to generate chemical substructures for MS/MS peak assignments in MetFrag**
*Christoph Ruttkies and Steffen Neumann* **112**

**P81 Gamification of gene prediction**
*Klas Hatje, Dominic Simm and Martin Kollmar* **113**

**P82 Identification of gene co-expression networks associated with different cellular and immunological states**
*Marc Bonin, Jekaterina Kokatjuhha, Stephan Flemming, Biljana Smiljanovic, Andreas Grützkau, Till Sörensen and Thomas Häupl* **114**

**P83 Combining features for protein interface prediction**
*Torsten Wierschin, Keyu Wang, Stephan Waack and Mario Stanke* **115**

**P84 Circular permutations: detecting evolutionary related protein pairs based on structure analysis**
*Martin Mosisch, Thomas Margraf and Andrew Torda* **116**

**P85 A scalable method for the correction of homopolymer errors**
*Giorgio Gonnella and Stefan Kurtz* **117**

**P86 Multiple genome comparison based on overlap regions of pairwise local alignments**
*Katharina Jahn, Henner Sudek and Jens Stoye* **118**

**P87 Enrichment Analysis for Hierarchical Clusters**
*Jan T Kim, Karen Staines, John Young, Zenon Minta, Krzysztof Smietanka, Devanand Balkissoon, Raul Ruiz-Hernandez and Colin Butter* **119**

**P88 A general approach for discriminative de-novo motif discovery from high-throughput data**
*Jan Grau, Stefan Posch, Ivo Grosse and Jens Keilwagen* **120**

**P89 Novel Visualization Approach Integrating Network and Structure Analysis of Proteins**
*Nadezhda T. Doncheva, John H. Morris, Eric F. Pettersen, Conrad C. Huang, Karsten Klein, Francisco S. Domingues, Thomas E. Ferrin and Mario Albrecht* **121**

**P90 Protein Subcellular Location Prediction Using Principal Component Analysis**
*Daichi Nogami, Yuichi Nakano and Yoshihiro Taguchi* **122**

**P91 Analyzing taxon and pathway coverage profiles with applications to metatranscriptomics**
*Daniela Beisser and Sven Rahmann* **123**

**P92 Bringing together only what belongs together: Characterizing and distinguishing protein structure families using distances based on contact map overlap**
*Inken Wohlers, Gunnar W. Klau and Rumen Andonov* **124**

# *Subti*Wiki – more than just a *Wiki*

Raphael Michna and Jörg Stülke
*Institute of General Microbiology, Georg-August University Göttingen*
rmichna@gwdg.de

*Bacillus subtilis* became important in many scopes of human life. The gram-positive and rod-shaped bacterium is used biotechnologically to produce antibiotics, vitamins and many more. Due to the fact that *B. subtilis* is closely related to the terrifying human pathogen *Staphylococcus aureus*, the investigation of new therapeutical methods becomes more and more important.

Therefore *Subti*Wiki[2] was initiated to collect the available information of the bacterial genom. Additionally information of products, regulation, references, mutants and so on were collected but at a certain point the information became overwhelming and the data was visualized in external applications. The first application, *Subti*Pathways[1], was created to present metabolic pathways and the genetic regulation. The second application, *Subt*Interact, was initiated to fix the enormous data of protein-protein interactions and the last one tries to unzip the information of a large microarray analysis where the transcript levels of all genes were measured under wide range of artificially created circumstances[3]. Finally *Subti*Wiki offers more than just a collection of genetic information. It tries to move forward to complete the knowledge on *B. subtilis* with the wiki as a base using several external applications.

We now try to combine the available applications to mix transcriptional with metabolic information using the GoogleMaps heatmap layer. The investigation should reveal new insights into the complex framework of *B. subtilis*. In the time of '*Omics*' the handling and the presentation of large datasets is more in focus than ever. Additionally, proteomics and metabolimocis data can be integrated to receive a more detailed overview on the organism.

# References:

**[1] Lammers CR, Flórez LA, Schmeisky AG et al.** Connecting parts with processes: *Subti*Wiki and *Subti*Pathways integrate gene and pathway annotation for *Bacillus subtilis*. Microbiology. 2010. 156: 849-859.

**[2] Mäder U, Schmeisky AG, Flórez LA et al.** *Subti*Wiki--a comprehensive community resource for the model organism *Bacillus subtilis*. Nucleic Acids Res. 2012. 40: D1278-D1287.

**[3] Nicolas P, Mäder U, Dervyn E et al.** Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis.* Science. 2012. 335: 1103-1106.

# Wondolin: A Web Service for Protein Sequence Comparison

**André Ahrens** and Martin Kollmar

*Group Systems Biology of Motor Proteins, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany*
anah@nmr.mpibpc.mpg.de

Comparing sequences is one of the basic tasks in bioinformatics. In most cases, this task is included in sequence search algorithms like FASTA, BLAST or BLAT, which aim to select and order sequences from a database by similarity to a query sequence. In these algorithms only two sequences are compared at once. Many sequences are compared in multiple-sequence-alignments (MSAa) that can be coloured according to amino acid conservation at each position. When working with long and many sequences, inspecting similarity via MSAs becomes very tedious for the user. The tool SimPlot visualizes similarity as line-plots, which is very useful for comparing long DNA sequences. However, the number of sequences is limited, and groups of sequences cannot be compared. Also, a tool that quantifies similarity between groups of sequences is missing. Here, we present a new web service for protein sequence comparison, which we called Wondolin. It allows to compare either an unlimited number of single sequences or to compare clusters of sequences. Clusters can be generated manually, by using CD-Hit or the UPGMA algorithm. UPGMA clusters can be computed based on sequence identity, sequence similarity or Kolmogorov complexity. Optionally, sequences/clusters can be compared locally, globally or internally (only for clusters). Depending on the selected options the similarity of selected sequence and clusters can be visualized as box-plots, line-plots of similarity or identity, conservation-coloured alignments, or similarity matrices. Wondolin is available at http://www.motorproteinde/wondolin.

# WaggaWagga: a Web Service to Compare and Visualize Coiled-Coil Predictions, and to Assess the Oligomerisation State of Coiled-Coils

**Dominic Simm**, Klas Hatje and Martin Kollmar

*Group Systems Biology of Motor Proteins, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany*
dosi@nmr.mpibpc.mpg.de

Coiled-coils belong to the most common structural motives for proteins. The sequences of coiled-coils are typically characterized by contiguous heptad repeats $(a\text{-}b\text{-}c\text{-}d\text{-}e\text{-}f\text{-}g)_n$ with hydrophobic residues in "*a*" and "*d*" positions and the remaining residues mainly charged thus favouring α-helical formation. In most cases, the individual α-helices are not very stable but become stabilized by wrapping around each other. Thus left-handed coiled-coils are formed, in which the hydrophobic residues are buried in the centre of the molecule. Coiled-coils can either be parallel, anti-parallel, homodimers, heterodimers, trimers, or any other oligomer. Several programs exist to predict coiled-coil regions like Marcoil, Multicoil or Paircoils. However, all these prediction programs are biased towards highly charged sequences. Charged residues are even found, although rarely, in "*a*" and "*d*" positions. Although predicted as homodimeric coiled-coils, many of these sequences in deed form stable single-α-helices instead. Here, we present a new web service for comparing and visualizing coiled-coil predictions, which we called WaggaWagga. The user can run Marcoil, Multicoil, Ncoils and Paircoils on the query sequence, and use Scorer 2.0, PrOCoil and LOGICOIL for oligomerisation prediction. The query sequence is visualized as helical wheel-diagram of parallel or anti-parallel homodimers, or parallel homotrimers, and as net diagram. With sliders the user can interactively move through the sequence automatically updating the visualizations. WaggaWagga is available at http://www.motorproteinde/waggawagga.

# OPTIMAS-DW: Integrating Different Maize –Omics Information into a Data Warehouse

Christian Colmsee[1], Tobias Czauderna[1], Anja Hartmann[1], Martin Mascher[1], Jinbo Chen[1], Matthias Lange[1], Falk Schreiber[1,2] and Uwe Scholz[1]

[1] *Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT Gatersleben, Corrensstr. 3, 06466 Stadt Seeland, Germany*

[2] *Martin Luther University Halle-Wittenberg, Institute of Computer Science, Von-Seckendorff-Platz 1, 06120 Halle, Germany*

colmsee@ipk-gatersleben.de

OPTIMAS-DW is a data warehouse integrating different plant –omics data including transcriptomics, metabolomics, ionomics, proteomics and phenomics data [CMC+12]. To enable users performing a cross-domain analysis, all stored data have to be linked together. Therefore, the data model includes metadata describing the experiments. Each measurement value can be connected to a specific sample, which has specific characteristics, such as genotype, plant growth stage, treatment or plant anatomy. In summary, there are data for 20 experiments with more than thirteen million measurement values public available. To link these data to other publicly available data resources, a LAILAPS search portlet is provided. LAILAPS is a life sciences search engine based on content-sensitive relevance ranking [LSC+10]. User profiles are used to perform a relevance prediction. Finally, user behavior is tracked to support the systems relevance factor estimation process. The LAILAPS portlet is integrated into the web interface of OPTIMAS-DW. When the user enters a search term or directly drags and drops a term from OPTIMAS-DW content into the search field, the user will be redirected to the LAILAPS search result page. Here, the user will find links to OPTIMAS-DW as well as to other resources such as Uniprot or Gene Ontology. With the described approach, OPTIMAS-DW enables the support of researchers within maize related research in general as well as with systems biological research in particular.

## References

[CMC+12] C. Colmsee, M. Mascher, T. Czauderna, A. Hartmann, U. Schlüter, N. Zellerhoff, J. Schmitz, A. Bräutigam, T.R. Pick, P. Alter, et al. OPTIMAS-DW: A comprehensive transcriptomics, metabolomics, ionomics, proteomics and phenomics data resource for maize. *BMC Plant Biology*, 12(1):e245, 2012.

[LSC+10] M. Lange, K. Spies, C. Colmsee, S. Flemming, M. Klapperstück, and U. Scholz. The LAILAPS search engine: A feature model for relevance ranking in life science databases. *Journal of Integrative Bioinformatics*, 7(3):118, 2010.

# Definitions and nomenclatures for alternative splicing events

Martin Pohl and Stefan Schuster
*Department of Bioinformatics, Friedrich Schiller University Jena*
m.pohl@uni-jena.de

Alternative splicing of pre-mRNAs in higher eukaryotes and several viruses is one major source of protein diversity.
A widely used graphical representation of alternative splicing events shows the alignments of transcripts as boxes representing exons connected by individual links for each isoform. Early in the analysis of alternative splicing, it became clear that standardization of the nomenclature for such alternative splicing graphs is important [Zav06]. Since then, some attempts have been made without leading to a broadly used and accepted nomenclature.
Here we give an overview of such notations. We revisit their suitability in terms of limitations to applicability, especially with respect to regulatory coupling of alternative splicing events. For example, there is no reason to exclude non-adjacent mutually exclusive exons. Hence, only approaches and nomenclatures considering mutual (perhaps long-ranging) dependencies within complete genes will have a chance of success in deciphering the full splicing picture.
We propose a general description to overcome identified limitations. It utilizes Boolean algebra reducing splicing data to basic information while still incorporating the full complexity of alternative splicing [Poh13].

## References

[Poh13] Martin Pohl, Ralf H. Bortfeldt, Konrad Grützmann and Stefan Schuster. Alternative splicing of mutually exclusive exons—A review, *BioSystems*, 114:31-38, 2013.

[Zav06] Mihaela Zavolan, Erik van Nimwegen. The types and prevalence of alternative splice forms. *Current Opinions in Structural Biology*, 16: 362–367, 2006.

# EndoNet: An information resource about the intercellular signaling network

Jürgen Dönitz and Edgar Wingender
*Institute of Bioinformatics, University Medical Center Göttingen*
juergen.doenitz@bioinf.med.uni-goettingen.de

The endocrine network plays a major role to transmit stimuli from the environment to the responsible targets and to coordinate cyclic events and the development of the organism. This intercellular network has an important impact on numerous metabolic and gene regulatory events in the scope of systems biology. In a first step, a source cell secretes a hormone. This messenger is transported to its target cells where it binds to specific receptors. The target cell has then two options to respond to the signal: The cell may adapt its phenotype or may be triggered to secrete the next hormone to build a signalling cascade.

EndoNet [1,2] is the first and only information resource that covers the components of the human endocrine network and their relations. Its content is annotated from scientific literature and each annotation references its source. The web interface provides interconnected detail pages to the main components of the endocrine system, i.e. hormones, receptors, cells and phenotypes, connected by entities like secretion, hormone binding or receptor-cell combination. Information for related anatomical structures are included with the help of the anatomical ontology Cytomer and the (ontology based answers (OBA) service [3,4]. The user can assemble its own network by choosing the complete network or selecting the nodes of interest. On the web pages the user can apply basic network analysis tools on the network or export it for further analysis in different file formats.

EndoNet is available at: http://endonet.bioinf.med.uni-goettingen.de

## References

[1] Dönitz J, Goemann B, Lizé M, Michael H, Sasse N, Wingender E, Potapov AP: EndoNet: an information resource about regulatory networks of cell-to-cell communication. *Nucleic Acids Res.* 2008, 36:D689-94.
[2] Potapov A, Liebich I, Dönitz J, Schwarzer K, Sasse N, Schoeps T, Crass T, Wingender E: EndoNet: an information resource about endocrine networks. *Nucleic Acids Res.* 2006, 34:D540-5.
[3] Cytomer: http://cytomer.bioinf.med.uni-goettingen.de
[4] OBA: http://www.bioinf.med.uni-goettingen.de/projects/oba

# BRENDA in 2013: integrated strains, kinetic data, improved disease classification

Placzek S., Chang A., Schomburg I., Schomburg D.
*Dpt. for Bioinformatics and Biochemistry, Technische Universität Braunschweig*
s.placzek@tu-braunschweig.de

BRENDA (BRaunschweig ENzyme DAtabase, http://www.brenda-enzymes.org) is the major database for enzyme functional information [1]. The manual collection of data from the primary literature and the curation of the database was started 25 years ago. Since then it has been continuously updated and further developed. The database covers many aspects of enzymology such as nomenclature, enzyme-catalyzed reactions, kinetic data, enzyme stability, purification, crystallization, or mutations. Each single data entry is connected to an organism name, to the literature reference and to the protein sequence identifier (if available). Organism-strain information are available for many data sets in the database. Thus it is possible to display data for e.g. Escherichia coli K-12 or Pseudomonas aeruginosa PAO1.

Since the huge amount of publications on enzyme properties does not allow the manual annotation of the complete literature of all enzymes, additional information is retrieved by textmining procedures based on text interpretation of sentences in abstracts and titles of the PubMed database: FRENDA and AMENDA [2] comprise organism-specific enzyme information, the latter including the enzyme source and the subcellular localization. DRENDA [3] connects enzymes with diseases. KENDA combines kinetic parameters of 13 categories with the enzyme- and organism-related terms found by FRENDA.

## References

[1] Schomburg I. et al. (2013). BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Research, 41, 764-772

[2] Chang A. et al. (2009). BRENDA, AMENDA, FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Research, 37, D588-D592

[3] Söhngen C. et al. (2011). Development of a classification scheme for disease-related enzyme information. BMC Bioinformatics, 12, 329

# Combining ontology design and flexible data management in biomedical sciences.

Timm Fitschen[1,2] Alexander Schlemmer[1,2] Daniel Hornung[1,2]
Philip Bittihn[1,2] Johannes Schröder-Schetelig[1,2] Ulrich Parlitz[1,2,3]
Stefan Luther[1,2,3]

*[1]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany [2]Institute for Nonlinear Dynamics, Georg-August-Universität, Göttingen, Germany [3]German Center for Cardiovascular Research (DZHK), partner site Göttingen, and Heart Research Center Göttingen, D-37077 Göttingen, Germany*
timm.fitschen@ds.mpg.de

Structured data management and retrieval remains to be a problem in cross-disciplinary scientific environments. Many existing approaches primarily involving semantic data representation are capable of implementing high structural complexity, but often require expert knowledge in ontology design, impose a rather static structure, and provide limited access for third-party software. Here we present HeartDB, a project aimed at the development of a database management system that fills the gap between non-object-oriented databases and ontology development tools. HeartDB provides a broad range of standards-compliant interfaces and programming language libraries allowing communication and data processing from third-party programs. The object-oriented structure of HeartDB is designed to ensure seamless integration of inhomogeneous data sources. A file server component manages references to file system objects providing transparent access to raw data files. The primary purpose is the management of multimodal experimental and simulation data from biomedical sciences, including biosignals and high-resolution fluorescence imaging data. However, other scientific purposes which require a flexible and dynamic structure are conceivable.

# Comparison of platforms to integrate transcriptome data from different sources

Sarah N. Mapelli [1,*], Bjoern Schumacher [2,*,¥], Ankit Arora [1,¥], Karsten R. Heidtke [1,*,¥]

[1] *ATLAS Biolabs GmbH , Friedrichstraße 147 , 10117 Berlin, Germany*

[2] *CECAD Cluster of Excellence in Ageing Research, University of Cologne, Germany*

mapelli@atlas-biolabs.com

**Background:** Along with the plethora of new opportunities in transcriptomic analyses generated by advances in high-throughput sequencing , several new artifacts and errors need to be interpreted . These are due to (i) technology dependent biases, (ii) analysis designs and (iii) the variability of processing strategies. A deeper understanding of similarities and discordances among different technologies is a basic requirement to detect and correct unknown systematic errors without prior biological knowledge of true positives and negatives.

**Methods:** Within the EU funded projects ITN CodeAge and aDDRess, transcriptome analyses of C.elegans were performed under 4 different experimental conditions on two different platforms: RNA-Seq using an Illumina HiSeq 2000 and expression profiling using a Whole Transcriptome Affymetrix microarray. Data analysis has been carried out with respect to genes and transcripts. After having defined a way to compare the different technologies, [IMN+12], their accordances and differences have been parsed. In particular, their performances have been compared in the detection of low-intensities expression patterns.

**Results:** Our analysis shows that in spite of the remarkable difference in dynamic ranges, resolutions and distributions, both results for raw expression intensities and differential expressions are highly correlating between the two platforms. Parsing intra-platform variance, it is possible for RNA-Seq to set an intensity threshold to optimize reliability. This is not applicable to arrays. The analysis of low expressions provides candidate lists with a highly dissimilar behaviour either respective to the same platform or compared with the other. These lists offered a basis to define patterns or features responsible for certain detection systematic failures.

**Conclusions:** Therefore, our study is aimed at defining guidelines for further analyses. We offer an additional way to interpret potential false positives and negatives in single platform studies and on the other hand we set the basis for integrating results derived from different platforms.

* *Marie Curie ITN CodeAge Project , Chronic DNA damage in Ageing*
*http://itn-codeage.uni-koeln.de/*
¥ *Marie Curie ITN aDDRess Project , Chromatin Dynamics on the DNA damage Response*
*http://www.itn-address.gr/*

## References

[IMN+12] Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, and Nielsen J. A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in saccharomyces cerevisiae. *Nucleic Acids Research*, Vol. 40, No. 20, 10084–10097 , 2012.

# Resource-Constrained Analysis of Ion Mobility Spectra with the Raspberry Pi

Elias Kuthe, Alexey Egorov, Alexander König, Marcel Köppen, Henning Kühn, Suzana Mitkovska, Marianna D'Addario, Dominik Kopczynzki and Sven Rahmann
*PG572 (project group "BreathBerry"), Computer Science XI, and Collaborative Research Center SFB 876, TU Dortmund, Germany*;
*Genome Informatics, Institute of Human Genetics, Faculty of Medicine, University of Duisburg-Essen, Essen, Germany*
Sven.Rahmann{at}uni-due.de

The analysis of the composition of exhaled breath has various applications in medical contexts. Ion mobility spectrometers with multi capillary columns (MCC/IMS) may yield valuable information about the presence of several diseases that would otherwise be hard to recognize. For the detection of volatile organic compounds (VOCs) by identifying peaks in measurement data, desktop-class computers are so far required to process the information gathered [HKM+13].

Our goal is to detect peaks in an online scenario (data stream) on resource constrained embedded hardware, the Raspberry Pi, eventually providing a single portable combined measurement and analysis device.

To identify peaks in an ongoing measurement, we designed a special Savitzky-Golay filter [SG64] which operates on a limited amount of single spectra. It identifies data points at the highest signal intensity of a given peak, computing a neighborhood of adjacent values most likely to belong to a single peak, eliminating the need for additional filtering before actively searching for peaks. Exploiting the filter's characteristics, we achieve sufficiently high performance to gain online capability on our target device.

## References

[HKM+13] A.-C. Hauschild, D. Kopczynski, D'Addario M, J. I. Baumbach, S. Rahmann, and J. Baumbach. Peak detection method evaluation for ion mobility spectrometry by using machine learning approaches. *Metabolites*, 3(2):277–293, 2013.

[SG64] A. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

# 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 [FK13]. 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 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.

[FK13] Frank Kramer, Tim Beißbarth, et al. rBiopaxParser - an R package to parse, modify and visualize BioPAX data. *Bioinformatics*, 2013 Feb;29(4)520-522.

# MarVis-Graph for integrative analysis of metabolic reaction chains in non–targeted experiments

Manuel Landesfeind, Alexander Kaever, Kirstin Feussner, Ivo Feussner, Peter Meinicke
*Georg-August-University, Institute of Microbiology & Genetics, Department of Bioinformatics, 37077 Göttingen, Germany*
manuel@gobics.de

Today, high–throughput technologies allow comprehensive analyses of organisms by combining data from different biological fields of study, such as metabolomics, transcriptomics, and proteomics. The development of algorithms and software which assist the integrated analysis of large datasets generated in these studies is a key challenge in current Bioinformatics.

The MarVis–Graph software provides a platform to investigate the full metabolism of an organism using combined data from non–targeted studies. Measurements from mass spectrometry, RNA–Seq or DNA micro–array are mapped to corresponding entities in metabolic networks compiled from the KEGG or BioCyc database collection. Reactions are scored based on the associated experimental data and, to cope with measurement errors, the Random–Walk–With–Restart graph algorithm [GBK$^+$12] is applied to refine the scores of nearby reactions. Afterwards, MarVis–Graph identifies sub–networks of connected high–scoring reactions which are ranked, evaluated with a random permutation test, and finally visualized for interactive inspection.

MarVis–Graph investigates full metabolic networks of organisms without restriction to separate pathways. The tool is particularly useful for discovery of reaction chains based on heterogeneous data from non–targeted experiments. Furthermore, MarVis–Graph may identify new pathways or connect existing pathways via known reactions. In combination with the MarVis–Suite [KLP$^+$12], MarVis–Graph has successfully been applied in wound response studies of Arabidopsis thaliana.

## References

[GBK+12] Enrico Glaab, Anas Baudot, Natalio Krasnogor, Reinhard Schneider, and Alfonso Valencia. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, 28(18):i451–i457, Sep 2012.

[KLP+12] Alexander Kaever, Manuel Landesfeind, Mareicke Possienke, Kirstin Feussner, Ivo Feussner, and Peter Meinicke. MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data. *Journal of Biomedicine and Biotechnology*, 2012.

# The MarVis-Suite: Integrative and explorative analysis of Metabolomics and Transcriptomics data

Alexander Kaever, Manuel Landesfeind, Kirstin Feussner, Ivo Feussner, Peter Meinicke
*Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-University Göttingen*
alex@gobics.de

The MarVis-Suite [KLP+12] combines the tools MarVis-Filter, MarVis-Cluster, and MarVis-Pathway for explorative analysis of large multivariate data sets from different omics platforms, such as mass spectrometry based Metabolomics or DNA microarray and RNA-seq based Transcriptomics. After data import from text files or Excel spreadsheats via MarVis-Filter, the data sets can be ranked and filtered by means of well-known statistical tests or highly customizable scoring methods, such as the signal-to-noise ratio. Filtered data sets can directly be combined by concatenating the features and merging the intensity/expression profiles. For a convenient overview and interactive cluster analysis, the profiles of the combined data set may be visualized via the MarVis-Cluster interface utilizing the robust training of one-dimensional self-organizing maps [MLK+08]. The features of filtered data sets or selected clusters may be annotated in the context of organism-specific pathways utilizing the MarVis-Pathway interface. Additionally, MarVis-Pathway provides an extensive statistical framework for a joint (Gene/Metabolite) Set Enrichment Analysis which utilizes the selection or ranking of data set features.

## References

[KLP+12] Alexander Kaever, Manuel Landesfeind, Mareike Possienke, Kirstin Feussner, Ivo Feussner, and Peter Meinicke. MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data. *Journal of Biomedicine and Biotechnology*, 2012, 2012.

[MLK+08] P. Meinicke, T. Lingner, A. Kaever, K. Feussner, C. Göbel, I. Feussner, P. Karlovsky, and B. Morgenstern. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms for Molecular Biology*, 3:9, 2008.

# FractalQC: A Bioconductor Package for Quality Control of RNA-Seq Coverage Patterns by Means of the Fractal Dimension

Stefanie Tauber and Arndt von Haeseler
*CIBIV - Center for Integrative Bioinformatics Vienna, MFPL*
stefanie.tauber@univie.ac.at

RNA-Seq is the state-of-the-art technology for global gene expression profiling. The initial research question is typically focused on differential expression inference for several biological sources. The analysis of RNA-Seq data is centered on the so-called count table, the number of reads per gene or exon. Preprocessing steps such as mapping and summarization are a mere means to generate this count table on which all subsequent statistical analyses base. Thus, by limiting the analysis to the count table, the unprecedented resolution of RNA-Seq, given by the per-base coverage pattern, is neglected.

In theory one would expect approximate uniform coverage along the gene regardless of the respective expression level. Yet, the observed coverage patterns may deviate greatly from this ideal due to biases which either stem from the library preparation in the wet-lab or from certain analysis strategies. Well known examples of induced biases include that random hexamer priming leads to a distorted nucleotide composition of the reads or that GC-poor as well as GC-rich genomic regions tend to have underrepresented read counts. In addition, outdated annotation or mapping heuristics may lead to wrongly placed reads and, as a consequence, to peculiar coverage patterns.

Here we argue not to be oblivious of the information contained in these patterns and propose a method, namely *fractalQC*, which identifies peculiar coverage patterns. *fractalQC* exploits the capability of the Fractal Dimension [Man82] to evaluate the roughness of a graph. By linking the roughness of the coverage graph to its reliability *fractalQC* unravels possible pitfalls while library preparation or analysis.

## References

[Man82] Mandelbrot,B.B. *The fractal geometry of nature.* W.H. Freeman and Co., San Fransisco,CA, 1982.

# Quantum Coupled Mutation Finder: Predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming

Mehmet Gültas, Martin Haubrock, Edgar Wingender and Stephan Waack
*Institute of Bioinformatics, Institute of Computer Science, University of Göttingen*
mehmet.gueltas@bioinf.med.uni-goettingen.de

The identification of functionally or structurally important non-conserved residue sites in MSAs is an important challenge for understanding the structural basis and molecular mechanism of protein functions. Despite the rich literature on compensatory mutations as well as sequence conservation analysis for the detection of those important residues, previous methods often rely on classical information-theoretic measures. However, these measures usually do not consider dis/similarities of amino acids which are likely to be crucial for those residues. In order to simultaneously incorporate significant similar and dissimilar amino acid pair signals in the prediction of functionally or structurally important sites, we develop the Quantum Coupled Mutation Finder (QCMF)[GWW13] applying the principles of quantum information theory. In QCMF, we present two major models based on quantum Jensen-Shannon divergence. The first model considers pairs of columns as an entangled quantum system and thus measures compensatory mutations between them. The second model provides a measurement for the sequence conservation of columns considering the pairs of columns as a separable quantum system. To demonstrate the effectiveness of our new method, we analyzed essential sites of two human proteins: epidermal growth factor receptor (EGFR) and glucokinase (GCK). The results show that the QCMF predicts a quite different set of residue sites as functionally and structurally important in comparison to previous CMF-method [GHTW12] and it reaches an improved performance in identifying essential sites of both proteins with a significantly higher Matthews correlation coefficient (MCC) value. On the top of that, results indicate that the residue sites found by QCMF are more sensible to catalytic sites, allosteric sites and binding sites than those found by the previous method.

## References

[GHTW12] Mehmet Gültas, Martin Haubrock, Nesrin Tüysüz, and Stephan Waack. Coupled Mutation Finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations. *BMC Bioinformatics*, 13(1):225, 2012.

[GWW13] Mehmet Gültas, Edgar Wingender, and Stephan Waack. Quantum Coupled Mutation Finder: Predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming. *BMC Bioinformatics,Under Review*, 2013.

# EnzymeDetector: Enhancements in 2013 for high throughput applications

Ulbrich M., Schomburg D.
*Dpt. For Bioinformatics and Biochemistry, Technische Universität Braunschweig*
m.ulbrich@tu-braunschweig.de

For reconstructing and analysing metabolic models an accurate annotation of enzyme functions is essential. Models of well studied organisms show inconsistencies and errors in existing databases. Therefore it is not recommended to rely solely on one source of annotations, but rather integrate and compare data of several databases. The EnzymeDetector [1] provides a fast up-to-date overview on sequence annotations improved by manually created organism-specific enzyme information from BRENDA [1], sequence pattern searches with BrEPS [2] and a sequence based similarity analysis by BLAST [3].

Further databases like KEGG orthology data [4] and PATRIC [5] are integrated as annotation sources. Improvements of the computational procedure, reliable database mapping, and enhancements on integrating all BRENDA annotation data and textmining data of AMENDA [1] according to reliability of results enriches EnzymeDetector's overall annotations and speeds up computing time.

Henceforth high throughput applications like discovering the whole enzyme stock of phylogenetic clades and comparisons of selected organisms by well supported annotations are done in workable time. Thereby clade-specific metabolic pathways may be recovered.

## References

[1] Schomburg et al. (2013). BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Research*, 41, 764.

[2] Bannert et al. (2010). BrEPS: a flexible and automatic protocol to compute enzyme-specific sequence profiles for functional annotation. *BMC Bioinformatics 2010*, 11, 589.

[3] Altschul et al. (1990). Basic local alignment search tool. *J. Mol. Biol.,* 215, 403-410.

[4] Kanehisa et al. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28, 27-30.

[5] Gillespie et al. (2011). PATRIC: The Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. *Infect. Immun*, 79, 4286-98.

# BAM
# A tool for basic analysis of microarray data

Marc Bonin, Florian Heyl, Irene Ziska, Lydia Ickler, Denise Sinning, Jekaterina Kokatjuhha, Thomas Häupl

*Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin*
marc.bonin@charite.de

**Introduction**
The definition of a standard of analyzing gene expression data concerning microarrays is crucial to get comparable results. Therefore a tool was developed with the help of the statistical language R. It can be downloaded for free from http://www.bamanalysis.charite-bioinformatik.de for local use or directly accessed online. Previous knowledge of R is not necessary. Furthermore the website provides documentation and guides to explain the application in detail which was integrated by Ruby, jQuery and JavaScript.

**Methods**
The tool is based on R version 2.15.3 including various packages from the Bioconductor repository which the R script loads in the beginning. The R script obtains the required CEL files from an input directory which has to be created manually if the tool is used locally. The results are saved in an automatically generated output directory. Using the online version the user will receive an e-mail containing a download link to the same files.

**Results**
First of all the R script normalizes the input data with a set of different methods like RMA, GCRMA, VSNRMA, MAS5. Afterwards quality control and RNA degradation are plotted to evaluate the sample quality. In addition a positive/negative distribution and a background intensity plot are generated to gain an insight of the hybridization quality. Besides box plots, histograms, MA plots and chip images are created. Therefore the different chips can be compared. Finally cluster dendrograms, PCA and correlation plots will approve previous assumptions. All in all the user will get a global overview of the analyzed data. Until now the tool supports the chips types HG U95A, HG U95A2, HG U133A, HG U133A2, HG U133AHT, HG U133ATAG, HG U133B, HG U133Plus2, HuGene 1, HG miRNA1, HG miRNA2, Mouse 430 2, Mouse 430A 2 and MG 74A 2.

**Conclusion**
In summary the main intention was to develop a tool defining a standard of analyzing microarrays which is simple to handle by everybody.

# DeNovoGUI: an open-source graphical user interface for de novo sequencing of tandem mass spectra

Thilo Muth, Lisa Weilnböck, Erdmann Rapp, Christian G. Huber, Lennart Martens,

Marc Vaudel and Harald Barsnes

Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg
muth@mpi-magdeburg.mpg.de

*De novo* sequencing is a popular technique in proteome bioinformatics for identifying peptides from tandem mass spectra (MS/MS) without having to rely on a protein sequence database. Moreover, *de novo* methods can be used to search for posttranslational modifications and mutated sequence variants. Despite their strong potential, the adoption threshold of *de novo* sequencing algorithms still remains quite high [All11]. We here present a user-friendly and lightweight graphical user interface called DeNovoGUI for running parallelized versions of the freely available *de novo* sequencing software PepNovo+ [FP05], greatly simplifying the use of *de novo* sequencing in proteomics. Our platform independent software is freely available under the permissible Apache2 open source license. Source code, binaries and additional documentation are available at http://denovogui.googlecode.com.

## References

[All11] Allmer J. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev Proteomics*, 8(5):645-57, 2011.

[FP05] Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling.. *Anal Chem*, 77(4):964-73, 2005.

# mascR: Efficient NGS fragment-size estimation

Orr Shomroni and Stefan Bonn
*German Center for Neurodegenerative Diseases*
Orr.Shomroni@dzne.de

NGS analysis pipelines require an accurate estimation of fragment size for short-end sequencing to improve on the performance of downstream analyses, such as peak calling and visualisation. Various algorithms were designed to estimate fragment sizes [Zhang08, Kharchenko08], but they show high variability in terms of finding the correct fragment size. The recent promising approach MaSC uses strand cross-correlation on uniquely mappable genomic regions to estimate the mean fragment size and was shown to reliably retrieve correct fragment sizes [Ramachandran13]. Here we present the R package mascR for shift-size estimation using an optimized variant of the MaSC algorithm. The package features a non-parametric reflective Pearson correlation as statistic, a speed-optimized algorithm, and allows for selecting small, highly informative genomic regions to reduce the algorithm runtime without reducing its performance. The package extends the R bit package and supports binary genomic mappability files for fast, random-access file operations. mascR is a standalone, speed-optimized tool for the rapid and reliable shift-size estimation of single-end NGS data that can be easily integrated into existing analysis pipelines.

## References

[Kharchenko08] Kharchenko, Peter V and Tolstorukov, Michael Y and Park, Peter J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* 26(12): 1351–9, 2008.

[Ramachandran13] Ramachandran, Parameswaran and Palidwor, Gareth A and Porter, Christopher J and Perkins, Theodore J. MaSC:

mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. *Bioinformatics*. 29(4): 444–50, 2013.

[Zhang08] Zhang, Yong and Liu, Tao and Meyer, Clifford A and Eeckhoute, Jérôme and Johnson, David S and Bernstein, Bradley E and Nusbaum, Chad and Myers, Richard M and Brown, Myles and Li, Wei and Liu, X Shirley. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*. 9(9): R137, 2008.

# New Network Analysis Tools Beyond Hairballs

Tim Kacprowski[1], Nadezhda T. Doncheva[2], Mario Albrecht[1,2]
[1]*University Medicine Greifswald, Greifswald, Germany*
[2]*Max Planck Institute for Informatics, Saarbrücken, Germany*

tim.kacprowski@uni-greifswald.de

A current challenge in bioinformatics is the development of methods and tools to integrate and analyze multiple heterogeneous datasets. Such datasets are often represented as biological networks. To understand the functional implications of biological networks, the incorporation of information about functional network modules is important. Since this information often lacks in network representations, we developed *ModuleGraph* [FoeK13], a Cytoscape plugin that supports the discovery, visualization, and analysis of functional modules. Furthermore, since many integrative prioritization methods of candidate disease genes and proteins rely on pre-defined data and offer the user only little control over the prioritization process, we also introduced *NetworkPrioritizer* [KaDA13], another Cytoscape plugin that alleviates these limitations. To rank candidates, it computes important topological centrality measures in both weighted and unweighted networks. Additionally, it provides state-of-the-art algorithms to compare and aggregate multiple rankings.

## References

[FoeK13] Sarah Foerster *et al.* Characterization of the EGFR interactome reveals associated protein complex networks and intracellular receptor dynamics. *Proteomics*, in press, 2013.

[KaDA13] Tim Kacprowski *et al*. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, 29(11):1471–1473, 2013.

# Identification of allele-specific expression and RNA editing sites in paired NGS data by ACCUSA2

Michael Piechotta[1] and Christoph Dieterich[2]
[1] *The Berlin Institute for Medical Systems Biology, MDC in Berlin*
michael.piechotta@mdc-berlin.de
[2] *Max Planck Institute for Biology of Ageing, Cologne*
christoph.dieterich@age.mpg.de

Direct comparisons of short read stacks are one way to identify Single Nucleotide Variants (SNVs) such as RNA editing sites or spots of allele-specific expression. SNV detection is specially challenging across samples with different read depths (e.g. RNA-Seq) and high background levels (e.g. selection experiments or RNA editing studies). We present ACCUSA2 to identify variant positions where nucleotide frequency spectra differ between two samples. To this end, ACCUSA2 integrates quality scores for base calling and employs a likelihood ratio test to identify variant sites. We performed read simulations of short read data on RNA editing and allele-specific expression scenarios and show that ACCUSA2 is superior to a state-of-the-art SNP caller. Our benchmarks implement read mixing scenarios (i.e. RNA editing scenario: one sample contains only reads that originate from a reference sequence and another sample contains reads from a mixture of a diverged sequence and the original reference). In the allele-specific expression scenario both read stacks contain a mixture of reads of the primary reference and a diverged sequence at random mixing ratios. We observed a superior sensitivity in the RNA editing and allele-specific expression scenarios (RNA editing benchmark: avg. sensitivity of 86.71% achieved compared to 72.30% of SAMtools/BCFtools) for ACCUSA2 while being comparable in SNP calling precision (on avg. 98.76% compared to 99.99%). We surveyed the performance of ACCUSA2 on publicly available data (`http://www.sanger.ac.uk/resources/mouse/genomes/`) from a large-scale RNA editing study of 15 mouse strains [DNkM$^+$12]. We assessed the accuracy of our predicted RNA-DNA differences (RDDs) by comparing them against a set of validated sites and contrasting our predictions with phylogenetic conservation of editing sites across multiple mouse strains.

## References

[DNkM+12] Petr Danecek, Christoffer Nellå ker, Rebecca E Mcintyre, Jorge E Buendia-buendia, Suzannah Bumpstead, Chris P Ponting, Jonathan Flint, Richard Durbin, Thomas M Keane, and David J Adams. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biology*, 13(4):26, 2012.

# Dynamics of Two-Photon Two-Color Transitions in Flurophores Excited by Femtosecond Laser Pulses

Peter S. Shternin[1], Andrey G. Smolin[1], Oleg S. Vasyutinskii[1]
Stefan Denicke[2], Sebastian Herbrich[2], Karl-Heinz Gericke[2]
[1]*Ioffe Institute, St.Petersburg, Russia*
[2]*TU Braunschweig, Braunschweig, Germany*
osv@pms.ioffe.ru

We present the results of theoretical and experimental studies of polarized fluorescence in flurophores excited by two-photon two-color (2P2C) laser pulses. Quantum mechanical expressions describing the fluorescence polarization have been derived under the condition of isotropic rotation diffusion for arbitrary polarization of each of the three photons involved in the photoprocess. The experiment has been carried out on p-terphenyl and 2-methyl-5-t-butyl-p-quaterphenyl (DMQ) dissolved in cyclohexane/paraffin. The fluorescence was produced within a 2C2P excitation scheme utilizing simultaneous absorption of two femtosecond laser pulses in the 400-440 nm and in the 800-880 nm spectral range. Using different combinations of the photon polarizations we extracted seven time-dependent molecular parameters from experiment. The analysis of the obtained experimental data was based on the *ab initio* calculations of the vertical excitation energies and transition matrix elements and allowed for determination of the whole structure of the two-photon absorption tensor, fluorescence lifetime, and rotational correlation times. This gives information on possible two-photon excitation channels and interaction of flurophores with surrounding solute molecules. The new technique developed will be used for determining bioinformation from fluorophores embedded in big biologically relevant molecules.

## References

[1] Peter S. Shternin, Karl-Heinz Gericke, Oleg S. Vasyutinskii. *Molecular Physics*, 108: 813-826, 2010.
[2] Stefan Denicke, Karl-Heinz Gericke, Andrey G. Smolin, Peter S. Shternin, Oleg S. Vasyutinskii. *J. Phys. Chem. A*, 114: 9681-9692, 2010.
[3] P. S. Shternin, A. G. Smolin, O. S. Vasyutinskii, S. Denicke, S. Herbrich, K.-H. Gericke. *Proc. SPIE*, 8553: 85531A, 2012.

# BiSQuID: Bisulfite Sequencing Quantification and Identification

Cassandra Falckenhayn, Günter Raddatz and Frank Lyko
*Division of Epigenetics, German Cancer Research Center, Heidelberg, Germany*
c.falckenhayn@dkfz-heidelberg.de

DNA methylation is a widely conserved modification and the focus of epigenetic research over the last few years [1]. Epigenetic modification patterns can be analyzed at single-base resolution using bisulfite-sequencing as a standard method [2]. The development of "next-generation" sequencing technologies, such as 454® and Illumina®, as well as the cost reduction resulted in an enormous sequencing depth for bisulfite-sequencing and the challenge of managing data analysis properly [3]. Several programs and web applications were designed to address this issue [4-7]. These tools have benefits and drawbacks, such as user friendly graphical surfaces and a limitation in the number of uploaded sequences, respectively. Here we present BiSQuID (**Bi**sulfite **S**equencing **Qu**antification and **ID**entification) as a tool which enables the user not only to analyze their bisulfite-sequencing data, but also to keep track of their experiments via a database. In contrast to other tools, the raw data of a whole sequencing run can be used as input. Based on the supplied primer and barcode sequences BiSQuID filters for reads of interest and thus different amplicons are analyzed in parallel. The results are visualized as heatmaps with a fixed size and can be saved as comprehensive packages. Moreover, via a downstream function of BiSQuID the results can be directly used for a differential methylation analysis.

## References

[1] Jaenisch R and Bird A 2003; *Nat Genet Suppl* 33:245-254
[2] Lister R and Ecker JR 2009; *Genome Res* 19(6):959-966
[3] Shendure J and Ji H 2008; *Nat Biotechnol* 26(10):1135-1145
[4] Rohde C *et al.* 2010; *BMC Bioinformatics* 11:230
[5] Kumaki Y *et al.* 2008; *Nucleic Acids Res* 36:W170-W175
[6] Gruntman E *et al.* 2008; *BMC Bioinformatics* 9:371
[7] Lutski P *et al.* 2011; *Nucleic Acids Res* 39:W551-W556

# Detection and monitoring of excited biomolecules by means of holographic technique

Irina Semenova, Oleg Vasyutinskii and Alexandra Moskovtseva
*Ioffe Physical Technical Institute, St.Petersburg, Russia*
irina.semen@mail.ioffe.ru

The major direct method being used to detect excited biomolecules involves the recording of a fluorescence signal induced by their radiative deactivation. However for many molecules of interest the radiative deactivation may be forbidden and thus the radiationless channel of deactivation predominates. In such cases indirect recording methods may be used being based on registration of thermal variations in a medium induced by the radiationless release of energy. One of the most commonly used methods is the thermal lens technique [BH92]. Although it allows one to reliably monitor temporal characteristics of a process, it does not provide information on spatial distribution of thermal disturbances. We suggest a novel approach based on the technique of holographic interferometry which allows one to obtain in a single shot a 2D image of the whole area under study. The recorded patterns provide data on spatial distribution of local variations of refractive index induced by temperature gradient. In our experiments the technique was applied to monitor the process of photosensitized generation and following radiationless deactivation of singlet oxygen molecules in water. The recorded holographic interferograms allowed us to reconstruct the temperature field in the area under study and thus to obtain information on the spatial distribution and dynamics of excited oxygen molecules. Being combined with fluorescence detection the holographic technique can provide a comprehensive data both on spatial distribution and temporal evolution of excited biomolecules in a medium.

## Reference

[BH92] Silvia E. Braslavsky' and George E. Heibel. Time-Resolved Photothermal and Photoacoustic Methods Applied to Photoinduced Processes in Solution. *Chem. Rev*., 92: 1381-1410, 1992.

# A spherical model of alveolar macrophages using computerized graphical techniques

Dominic Swarat[1], Martin Wiemann[2] and Hans-Gerd Lipinski[1]
[1] *University of Applied Sciences and Arts, Dortmund*
[2] *IBE R&D Institute for Lung Health gGmbH, Münster*
swarat@gmx.de

Alveolar macrophages enclose particles within membrane bound vacuoles. While changing their form from round to flat [DBA79], the amount of engulfed particles may be limited by the cell's volume [Mor92] or by the total area of its membrane. Here a computerized graphic model was developed to predict the maximum number of nano- or micron-sized ingestible particles. A round sphere was taken as a basic geometrical form. We created a computer program which allowed us to stretch or compress this basic form under stereoscopic control to form a more realistic macrophage model. As the real cell's nucleus deforms the upper surface, the geometry of the model was adapted accordingly. Utilizing information about size, form, and shape of macrophages from phase contrast images, confocal laser scanning, and electron microscopy, the cell model could be compared with real cells. In its current state the computerized graphic cell model was useful to dynamically rebuild the form of a real macrophage. Thereby, surface and volume of the modeled cell could be evaluated with typical geometry parameters, which is impossible under real microscopic conditions. If volume and surface of a real macrophage were implemented into the model, the uptake into phagosomes of user-defined particle collectives could be modeled. Limitations were reached either by volume or cell membrane settings.

## Acknowledgement

This work was funded by the German Ministry of Education and Research (BMBF / Frderkennzeichen 17PNT026)

## References

[DBA79] Gerald S. Davis, Arnold R. Brody, and Kenneth B. Adler. Functional and Physiologic Correlates of Human Alveolar Macrophage Cell Shape and Surface Morphology. *Chest*, 75(2):280–282, 1979.

[Mor92] P.E. Morrow. Dust Overloading Issues in Toxicology of the Lungs: Update and Appraisal. *Toxicology and Applied Pharmacology*, 113(1):1–12, 1992.

# Computing metabolic costs of amino acid and protein production in *Escherichia coli*

Christoph Kaleta[1], Sascha Schäuble[1], Ursula Rinas[2,3] and Stefan Schuster[4]

[1]*Research Group Theoretical Systems Biology, Friedrich Schiller University Jena, Jena, Germany*
[2]*Helmholtz Centre for Infection Research, Braunschweig, Germany*
[3]*Institute of Technical Chemistry - Life Science, Leibniz University of Hannover, Hannover, Germany*
[4]*Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Pl. 2, 07743 Jena, Germany*
Stefan.schu@uni-jena.de

*Escherichia coli* is the most widely used microorganism for the production of recombinant proteins and is of increasing importance for the production of low-molecular weight compounds such as amino acids. The metabolic cost associated with the production of amino acids and (recombinant) proteins from the substrates glucose, glycerol and acetate was determined using three different computational techniques, including linear programming. That allowed us to identify those amino acids that put the highest burden on the biosynthetic machinery of *E. coli*. Comparing the costs of individual amino acids, we found that methionine is the most expensive amino acid in terms of moles of ATP consumed per mole produced while leucine is the most expensive amino acid when additionally cellular abundances of amino acids are taken into account [1]. Moreover, we show that the biosynthesis of a large number of amino acids from glucose and particularly from glycerol provides a surplus of energy, which can be used to balance the high energetic cost of amino acid polymerization [1]. Our results are based on systematic calculations and, thus, considerably improve earlier, practically manual calculations [2,3].

## References

[1] C. Kaleta, S. Schäuble, U. Rinas, S. Schuster, *Biotechnol. J.* (2013) June 7 Epub ahead of print, doi: 10.1002/biot.201200267.
[2] A.H. Stouthamer, *Antonie Van Leeuwenhoek* (1973) 39, 545–565.
[3] H. Akashi, T. Gojobori, *Proc. Natl. Acad. Sci.* (2002) USA 99, 3695–3700.

# The HGT Calculator: targeted detection of horizontal gene transfer from prokaryotes to protozoa in small data sets

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

Detection of horizontal gene transfer (HGT) is usually based on search for atypical sequences in a genome of interest. Computational methods are mostly designed for genome-wide detection. Often these methods result in a list of ancient endosymbiotic gene transfer events (EGT) or a lot of uncharacterized proteins after scanning whole genomes. Each method has its advantages and disadvantages, and under certain conditions they are inconsistent with each other. Often a method is also restricted to a special field of research and cannot be easily applied to another area. Thus, it is difficult to find a suitable method for prokaryote-to-eukaryote transfer.

We intended to scale down the effort to find HGT events, to employ simple algorithms and incomplete data sets. We were interested in more recent events of specific gene acquisition in protozoa, which had occurred after the separation of defined species. How can we search for HGT, if the genome sequence of the organism under study is not completely available or if we are interested in the evolution of a specific gene, not a specific organism? Is it possible to detect single HGT events between prokaryotes and eukaryotes in small data sets, like they are common in research practice, even without knowing the corresponding species tree, which would be needed for some HGT detection methods?

For this purpose we developed the HGT Calculator, a tool combining four different tree based and non-tree based approaches for the detection of HGT in a score-based application. We analyze protein domains, search for a high ratio of prokaryotic BLAST hits, untypical GC content, a low codon adaptation index, and present a new approach for multi-level analysis of phylogenetic trees.

Here, we show the potential of the program on the example of isocitrate dehydrogenases in the parasitic protozoon *Leishmania major*.

# Linking Phenotypes and Genomic regions: the Forward Genomics Approach

Michael Hiller and Xavier Prudent
*Max Planck Institute of Molecular Cell Biology and Genetics, Dresden,*
*Max-Planck-Institut für Physik komplexer Systeme, Dresden*
hiller@mpi-cbg.de, prudent@mpi-cbg.de

Evolution has produced a great phenotype diversity among all species. Today we have many sequenced genomes but it mostly remains unclear, which genomic regions are responsible for phenotype differences. Previously, we developed a forward genomics method to match the divergence pattern of conserved genomic regions to the loss pattern of a given phenotype across a phylogeny, focusing on independent phenotype losses [ea12][HMG12]. The original forward genomics method relied on finding a genomic region that is more diverged in all species where the phenotype is lost. While this simple approach works for some phenotypic differences [ea12], it leaves room for improvement. Specifically, this simple approach only indirectly uses the fact that independent species have lost the same phenotype and it completely ignores that species evolve at different speeds, which affects the divergence of genomic regions. Here, we present and compare new methods to detect genomic regions responsible for phenotype differences in a proper statistical framework. Our new approaches take the evolutionary relatedness of the considered species (phylogeny) as well as their molecular evolutionary rates into account. We use simulations that evolve entire genomes in silico to generate test sets under realistic parameters and present a performance comparison of these methods on these test sets.

## References

[ea12] Hiller M. et al. A forward genomics approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Reports*, 2(4):817–823, 2012.

[HMG12] Schaar BT. Hiller M. and Bejerano G. Hundreds of conserved noncoding genomic regions are independently lost in mammals. *Nucleic Acids Res*, 40(22):1146311476, 2012.

# Automated combined analysis of DNA methylation and transcription profiles in different immune cells

Marc Bonin[1], Lorette Weidel[1], Stephan Flemming[2], Andreas Grützkau[3], Biljana Smiljanovic[1], Till Sörensen[1], Stefan Günther[2], Thomas Häupl[1]

[1]*Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin*
[2] *Institute of Pharmaceutical Sciences, University of Freiburg*
[3] *German Arthritis Research Center, Berlin*
marc.bonin@charite.de

**Introduction**
Site specific methylation of DNA may contribute to the regulation of gene expression. To test for functional CpG sites, we compared methylation with transcription in parallel in different sorted immune cell types. In order to perform primary analysis and to map corresponding results, a software tool was needed.

**Methods**
Cells from 4 healthy donors were sorted by FACS technology for naive and activated/memory T-cells and B-cells, NK-cells, monocytes, and granulocytes. Genome wide DNA methylation was assessed using the HumanMethylation450 BeadChip platform and Genome Studio (Illumina). Transcriptomes were determined with Affymetrix HG-U133 Plus 2.0 GeneChips. A tool has been implemented in Java and R to import and analyse transcription and methylation data, to determine high and low transcribed genes, to match them with the status of DNA methylation and to save the results as .txt and .jpg files. The tool will be provided on our homepage http://www.charite-bioinformatik.de. First tests of the tool were done with T-cells and monocytes.

**Results**
Differences of gene expression or DNA methylation were higher between more distant cell types like monocytes and T-cells (4624 genes; 19261 sites) and lower between closer related cells like naive and activated/memory cells of the same lymphocyte subtype (CD4+ T-cells: 638 genes; 9412 sites). Comparing monocytes against T-cells, corresponding changes of expression and methylation were found in only 629 (279) of 1951 increased (2673 decreased) expressed genes. Comparing methylation between memory and naive T-cells revealed a shift especially at sites with high methylation in naive to less methylation in memory cells. Nevertheless, corresponding changes of expression and methylation were similar for increased (57 of 332) and decreased (53 of 306) expression in memory versus naive T-cells, suggesting unspecific differences of methylation possibly related to proliferation of activated/memory cells. Of all CpG sites on the BeadChip, which were annotated to an individual of these identified genes, only about 10% were concordant with expression.

**Conclusion**
Corresponding information of transcription and methylation is indispensible to infer methylation associated gene regulation. This applies not only for microarray but also for sequencing approaches, as the methylation status seems

# Automated Classification of Cell Populations with Multi-channel Flow Cytometry Data - Using Sparse Grids Classifying A Sparsely Populated Data Space

Manuel M Nietert[1]*, Steve Wagner[2], Annalen Bleckmann[1], Klaus Jung[1], Andreas Schneeweiss[3], Dorit Arlt[2,4], Tim Beißbarth[1]

*1 University Medical Center Göttingen, Department of Medical Statistics, Humboldtallee 32, 37073 Göttingen, Germany*
*2 German Cancer Research Center (DKFZ), Division Stem Cells and Cancer, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany*
*3 National Center for Tumor Diseases (NCT) Heidelberg, Im Neuenheimer Feld 460, 69120 Heidelberg, Germany*
*4 University Medical Centre Freiburg, Centre of Chronic Immunodeficiency , Breisacher Straße 117 , D-79106 Freiburg i. Brsg*
manuel.nietert@med.uni-goettingen.de

Assigning cell population clusters to multi-channel FACS data is a routine task in diagnostic settings. An automated strategy based on analyzing the build-up of reference sets can be used to generate classification models helpful in assigning similar classes to the various populations present in new sets. For most of the available experimental FACS data, each resulting data space of an experiment is though only sparsely populated. This fact has various reasons ranging from the currently observed populations in the set likely not spreading out over all available combinations of measured attributes to an incomplete representative sampling of the population distribution itself. By applying a sparse grid-based approach to classify the multi-dimensional FACS space derived from patients blood samples we present a means to automatically assign cell populations based on previously defined reference populations, while in this case aiding in the identification of potential circulating tumor cells (CTCs) minimizing the prior use of expert knowledge, whilst still optimizing the sensitivity and specificity of the classification method.

# Spatial distribution of cells in Hodgkin Lymphoma

Tim Schäfer, Hendrik Schäfer, Jörg Ackermann, Norbert Dichter, Claudia Döring, Sylvia Hartmann, Martin-Leo Hansmann, Ina Koch

*Molecular Bioinformatics, Institute of Computer Science, Cluster of Excellence "Macromolecular Complexes", Johann Wolfgang Goethe-*

*University Frankfurt am Main, Robert-Mayer-Straße 11-15, 60325 Frankfurt am Main, Germany*

*Senckenberg Institute of Pathology, Johann Wolfgang Goethe-University Frankfurt Main, 60590 Frankfurt am Main, Germany*

Hodgkin lymphoma (HL) is a type of B cell lymphoma which arises from germinal center B cells. HL differs from other lymphoma and cancer types in the morphology and distribution of the malignant cells. Most notably, no solid tumour is formed, and the malignant HRS cells make up only a small fraction of the cells in the affected tissue. Modern immunostaining protocols and the acquisition of high-resolution images for diagnostic purposes in pathological labs currently lead to large and growing databases of HL images. Exploring these images using automated image analysis is a hard task but may lead to a deeper understanding of HL. Here we present our current work on the analysis of a database of HL whole slide images. We performed pre-processing and identified regions of interest (ROI) that contain CD30-positive cells in the images as described before. A CellProfiler pipeline was used to detect and measure cells in the ROI. The cells were then classified using shape descriptors and stored in a database with their 2D coordinates. This allowed for a statistical analysis of the spatial distribution of HRS cells within the tissue.

# Sequence based analysis of plant myosins

Stefanie Mühlhausen and Martin Kollmar

*Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for biophysical Chemistry, Göttingen*
stmu@nmr.mpibpc.mpg.de

The evolution of land plants is characterized by whole genome duplications, which drove species diversification and evolutionary novelties. Detecting these events is especially difficult if they date back to the origin of the plant kingdom. Established methods for reconstructing whole genome duplications include phylogenetic tree constructions, KS age distribution analyses, and intra- and inter-genome comparisons to detect synthenic regions.

778 myosins were identified in 67 sequenced plant genomes and their sequences assembled manually. Phylogenetic trees of 687 complete myosin motor domains revealed subfamily relationships and were consistent with recent species trees. The orthologous and paralogous relationships between sequences were incorporated into a new general and extendable naming scheme for plant class VIII and class XI myosins. Based on gene structure and protein sequence comparisons consensus domain architecture schemes were developed for both types. We specified motifs for the newly defined N-terminal MyTH8 domain of class VIII myosins and reassessed the definition of the DIL domain. Most of the myosins follow these consensus schemes except for three class XI myosin subtypes one of which is a headless myosin and another represents a short-tailed myosin missing the DIL domain. Liliopsida contain specific class XI myosins that are characterised by a conserved 800 residues insertion coding for and extending the coiled-coil region. Based on the myosin inventories and the phylogenetic tree of the motor domains 23 whole genome duplications have been reconstructed in plant evolution.

We could show that myosins are very suited to reconstruct whole genome duplications in plants. In contrast to other protein families, many myosins are still present in extant plants, they are closely related and have similar domain architectures, and their phylogenetic grouping follow the genome duplications. Because of the broad taxonomic sampling the dataset provides the basis for reliable future identification of further whole genome duplications in so far uncovered subbranches.

# Transcriptome analysis of the model organism *Tribolium castaneum*

Sarah Behrens, Robert Peuß, Barbara Milutinovic, Hendrik Eggert, Daniela Esser, Philip Rosenstiel, Erich Bornberg-Bauer, Joachim Kurtz
*Institute for Evolution and Biodiversity, University of Münster*
sbehrens@uni-muenster.de

*Bacillus thuringiensis is a natural microparasite of the red flour beetle Tribolium castaneum. In order to characterize the transcriptomic response in the host Tribolium castaneum upon infection with Bacillus thuringiensis, we have performed RNA-seq experiments for two different host populations - a commonly used laboratory strain and a newly collected, genetically diverse field population - at different time points after infection using two different methods of infection. After RNA isolation, in total 48 samples were sequenced using the Illumina HiSeq 2000 Sequencing System yielding a total number of 2.9 billion paired-end reads. We have established an NGS pipeline to detect and functionally characterize differentially expressed genes and to visualize gene expression differences between different host populations and infection methods. As a result, we found e.g. a higher number of differentially expressed genes in the genetically diverse field population compared to the commonly used laboratory strain for both infection methods. Furthermore, gene expression profiles differ strongly between the two infection methods, indicating that immunological and physiological processes underpinning the different routes of infection are clearly distinct. This project is part of the DFG priority programme "Host parasite coevolution".*

# Analysis of Wt1 ChIP Seq data from mouse glomeruli

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

The Wilms tumor suppressor gene 1 (*Wt1*) is known for its role in urogenital development. *Wt1* encodes a zinc finger transcription factor, which exists in two major splice variants due to insertion or deletion of the tripeptide KTS located between zinc finger III and IV. In adult kidneys, *Wt1* expression is restricted to podocytes, which are cells that keep the integrity of glomerular filtration barrier and contribute to the maintenance of kidney homeostasis. Podocyte-sepecific *Wt1* deletion in adult kidney of mouse result in focal segmental glomerulosclerosis (FSGS) [GKD+13].

However, the *Wt1* controlling transcriptional network remains unclear. We established the Wt1 ChIP-Seq method from freshly prepared wild type glomeruli to identify possible Wt1 binding sites in the genome and predict *Wt1* direct target genes in podocytes. Therefore, ChIP-Seq reads were aligned to the genome and peaks were called using MACS [FLQ+12]. This information was used to predict a binding motif for Wt1 using MEME [BBB+09] and to generate a list of possible target genes. This list contains already known targets like *nphs1* or *podx*, but also new target genes. In comparison to microarray data from *Wt1* knockdown mice, we could reveal new insights for regulatory functions of *Wt1*.

## References

[BBB+09] Timothy L Bailey, Mikael Boden, Fabian a Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37, July 2009.

[FLQ+12] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying ChIP-seq enrichment using MACS. *Nature protocols*, 7(9):1728–40, September 2012.

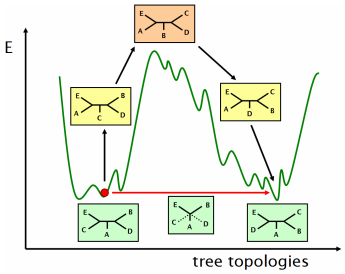
[GKD+13] Christoph A Gebeshuber, Christoph Kornauth, Lihua Dong, Ralph Sierig, Christoph Englert, Javier Martinez, and Dontscho Kerjaschki. Focal segmental glomerulosclerosis is induced by microRNA-193a and its downregulation of WT1. *Nature medicine*, 19(4):481–7, March 2013.

# Newtonian dynamics in the space of phylogenetic trees

Björn Hansen and Andrew E. Torda
*Center for Bioinformatics, University of Hamburg, Germany*
hansen@zbh.uni-hamburg.de

A classic unrooted phylogenic tree is a simple undirected graph. Edges are either present or absent and searching for a phylogenetic tree is a discrete optimisation problem. We have been developing an alternative view. Allowed trees are just points within a continuous space. Connections are continuous properties which behave like coordinates. If we know the similarities between objects like sequences, we can see how well the set of coordinates (connections) fits the experimental data. The greater the disagreement, the greater the force acting on the connections. This leads to a method for generating phylogenetic trees. One can perform classic, conservative Newtonian dynamics in the space which includes all possible trees.

A problem with current methods is that there may be more than one tree that is supported by the data. If they are separated by high barriers, as in the figure, current sampling methods will find it difficult to reach both trees [MV05].




At the moment, we are limited to distance-based phylogenies, but the method has advantage over Monte Carlo methods, that it uses gradient information, so sampling can be quite efficient. Like Monte Carlo methods, extensions such as simulated annealing or replica exchange are easy to implement. We see the long term benefit as a means of providing efficient sampling for seeding more sophisticated methods such as Bayesian inference.

## References

[MV05] E. Mossel and E. Vigoda (2005), Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees, *Sience*, 309:2207-9.

# Design of new inhibitors for HIV-Integrase: Implications of structure based drug design by Molecular Modeling approach

J. K. Gupta , K.P Nandhini, B. A. Cynthia , T. G. Krishnan , S.A.H. Naqvi

*OSDD Research Unit, Indian Institute Of Science, Bangalore, India. Department of Biotechnology, Sathyabama University, Chennai, India. Department of Biotechnology, Anna University, Chennai, India. BioDiscovery Group- Solutions for Future, India*

*contactme.asif@gmail.com.*

HIV-1 Integrase is a clinically validated therapeutic target for the treatment of HIV-1 infection, with one approved therapeutic currently on the market. This enzyme represents an attractive target for the development of new inhibitors to HIV-1 that are effective against the current resistance mutations. The enzyme consists of three domains. The N-terminal domain has a novel zinc-binding fold, the catalytic domain shares a common structural motif with other polynucleotidyl transferases, and the C-terminal DNA-binding domain has a Src-homology-3-like fold. This structural information provides the basis for drug development. To elucidate the structural properties of Coumarin a compound reported as anti-HIV (L. W. Law; Cancer Res., 1, 397 (1941) important for anti HIV activity, molecular modeling study was done on a set of 3000 2H-chromen-2-one derivatives which might selectively inhibit the strand transfer reaction of the HIV-1 Integrase. The docking result of the study of 3000 molecules demonstrated that the binding energies were in the range of -9.59 kcal/mol to -2.79 kcal/mol, with the minimum binding energy of -9.59 kcal/mol. We report molecule JK-1, JK-2, JK-3, JK-4 and JK-5, which showed H- Bonds with binding pocket of HIV-1 Integrase and promising ADMET results. 2 The molecule JK-1 showed Drug Likeness score of -0.79 with Mol PSA as 72.48 A and MolVol as 205.75 3 A . The MolLogS was -1.99 (in Log(moles/L)) 2493.00 (in mg/L). The LD 50 calculated for Rat/Intraperitoneal as 0.14 and Rat/Oral as 0.34. The predicted value for Human Bioavailability came out to be %F(Oral) > 30%: 0.033 and probabilities of effect on Blood was 0.77, Cardiovascular System 0.42, Gastrointestinal Systems 0.55, Kidney 0.07, Liver 0.14 and Lungs 0.37. The molecule JK-5 showed better results of Drug Likeness as -0.48 and probabilities of effect on Blood was 0.32, Cardiovascular System 0.27, Gastrointestinal Systems 0.09, Kidney 0.04, Liver 0.12 and Lungs 0.14. Further in-vitro and in-vivo study is required on these two molecules to design new derivatives with higher potency and specificity.

# Elucidating soil microbial communities in agricultural soils

Yudai Suzuki[1], Kazunari Yokoyama[2], Naomi Sakuramoto[3], Y-h. Taguchi[4]
[1]*Department of Phyics Chuo University*,[2]*National Agricultural Research Center*, [3]*DGC Technology Inc.* [4]*Department of Phyics Chuo University*
[1]nobitakun0ten@yahoo.co.jp, [2]kazunari@affrc.go.jp, [3]sakura@dgc.co.jp, [4]tag@granular.com

Suppressing soil diseases greatly improves the stability and sustainability of food production. The crisis in the ability of Japan to be self-sufficient in food illustrates the importance of improving the sustainability of food supplies. Two researchers recently applied numerical models and multivariate analysis to carbon resource consumption rates of soils measured using the OmniLog PM system[YT13]. However, direct evidence for interactions between microorganisms has not yet been obtained. We attempted to find evidence for interactions between microorganisms using the Markov Chain Monte Carlo (MCMC) method, assuming that a LotkaVolterra type model for interactions between soil microorganisms was valid. We made use of soils provided by a number of domestic companies and research institutes in this experiment. The sample of sick soil and the healthy soil was diluted, and put on 95 kinds of specified sources of carbon. A coloring pattern that exhibits carbon-resource consumption was evaluated every 15 minutes by a OmniLog PM system, and samples are carried out for a total of 48 hours. In the case of healthy soil, compared with sick soil, it was clear that there are many amounts of carbon-resource consumption in this result. In this research, using the above-mentioned experimental data, the increase of microbe and carbon-resource consumption were assumed to follow a Lotka-Volterra equation. We required result coinciding experimental data and a calculation data by MCMC method. Furthermore, the feature of each microbial community can be acquired by the source predation degree of carbon-resource consumption from computation.Disease suppressive soils turned out to have more microbiological biodiversity.

## References

[YT13] Kazunari Yokoyama,Y-h. Taguchi, Microbiology- and Biodiversity-Based Modeling of Suppression of Cottony Leak of Scarlet Runner Bean in Soils with Diverse and Uniform Ecology, Journal of Agricultural Science and Applications, vol.2, No.1, PP.35-44, (2013).

# A Novel Approach for Determining Spatial Colocalization of Proteins Inside Ceramide-rich Domains

Christian Imhäuser[1], Heike Gulbins[2], Erich Gulbins[2] and Hans-Gerd Lipinski[1]
[1]*Biomedical Imaging Group, University of Applied Sciences and Arts, Dortmund, Germany*
[2]*Institute for Molecular Biology, University Essen-Duisburg, Essen, Germany*
chrisdedisk@alice-dsl.net

Aggregation of receptor and signalling molecules (such as CD95 or CD40) within ceramide-enriched membrane domains results in a very high density of these proteins facilitating activation of associated enzymes, the exclusion of inhibitory molecules and/or the recruitment of further signalling molecules to transmit the signal into the cells. We could previously show that the hydrolysis of sphingomyelin by the activity of the acid sphingomyelinase leads to the formation of ceramide-enriched membrane platforms that serve the trapping and clustering of activated receptor molecules. [GJRea01] Ceramide-enriched domains were exhibited to be required for the induction of cell death by CD95, DR5, radiation, UV-light, and the infection of mammalian cells with some pathogenic bacteria, for instance P. aeruginosa. [GJRea03] [KP01]
However, at present the exact mechanisms of receptor clustering and distribution of proteins within the ceramide-enriched domains are unknown. For that reason, we generated digital images from anti-CD95 stimulated JY-cells that were stained with FITC-coupled anti-ceramide and Cy3-labelled anti-CD95 antibodies. We developed and adapted image filtering methods enhancing contrast, reducing noise caused by sensor sensitivity, and scaling logarithmically the obtained spatial image data due to supposed unspecified adherences of fluorescent antibodies. From these optimized image data we generated 3D models of JY-cells using adjusted volume and surface reconstruction algorithms. To detect colocalizations of CD95 and ceramide molecules we created several different computerized methods for qualitative and quantitive analysis. Using rasterizing 3D data of each channel into spatial cells, detecting relevant intensity values, and computing respective colocalization values we could determine the degree of colocalization. Spatial fluorograms, re-association and visual-

ization of clipped areas inside fluorograms allowed for interactive analysis of colocalization domains. A novel developed alternative method in addition to well-established techniques allowed for a quantitative analysis of the spatial arrangement of proteins in ceramide-rich domains of living cells.

## Acknowledgement

This work was funded by the German Ministry of Education and Research (BMBF / Förderkennzeichen 17PNT026).

## References

[GJRea01] H. Grassmé, A. Jekle, A. Riehle, and et al. CD95 Signaling via Ceramide-rich Membrane Rafts. *J. Biol. Chem.*, 276(23):20589–20596, 2001.

[GJRea03] H. Grassmé, V. Jendrossek, A. Riehle, and et al. Host Defense against Pseudomonas Aeruginosa Requires Ceramid-rich Membrane Rafts. *Nat. Med.*, 9(3):322–330, 2003.

[KP01] C. Krawczyk and J. M. Penninger. Molecular Controls of Antigen Receptor Clustering and Autoimmunity. *Trends Cell. Biol.*, 11(5):212–220, 2001.

# Structure modeling of proteins for the biosynthesis of sex pheromones in zygomycetous fungi

Sabrina Ellenberger and Johannes Wöstemeyer
*Chair of General Microbiology and Microbial Genetics, Friedrich Schiller University Jena, 07743 Jena, Neugasse 24, Germany*

4-Dihydromethyltrisporate dehydrogenase (TSP1, [1]) is a NADP-dependent oxidoreductase acting on trisporoids. This gene is part of a pheromone biosynthesis pathway in zygomycetous fungi. Trisporoids are important for recognition of complementary mating types and sexual spore formation. They are synthesized by oxidative degradation of $\beta$-carotene. Recently, we presented a model for the tertiary structure of TSP1 in *Mucor mucedo*. Xylose reductase from the yeast *Candida tenuis* turned out to be a suitable template for modeling this protein. We were interested in protein structures of TSP1 from other zygomycetes and the deciding differences between xylose reductases and TSP1.

Genes, putatively encoding the TSP1-dehydrogenase, cannot directly be detected by simple sequence similarity approaches. We wanted to investigate, if it is possible to find TSP1 in other fungi by a text-based search for xylose reductase or via protein domains for aldo-keto reductases in JGI Genome Portal MycoCosm. To check if the detected sequences were TSP1 genes, we analyzed them by modeling of tertiary structures with the Phyre2 server [2] and docking approaches with AutoDock [3]. We compared the results with the former analysis of TSP1 from *Mucor mucedo* and found sequences for *Phycomyces blakesleeanus*, *Rhizopus oryzae* and *Umbelopsis ramanniana*. The docking studies revealed additional information on substrate specificity of the different organisms for special trisporoid derivatives.

## References

[1] K. Czempinski, V. Kruft, J. Wöstemeyer, and A. 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] L.A. Kelley and M.J.E. Sternberg. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.*, 4:363–371, 2009.

[3] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, and A.J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30:2785–2791, 2009.

# Functional and metabolic characterization of plant peroxisomal proteomes

Ana Tzvetkova, Sigrun Reumann, Peter Meinicke and Thomas Lingner
*Department of Bioinformatics, Institute for Microbiology and Genetics, University of Göttingen, Göttingen, Germany*
thomas@gobics.de

Peroxisomes are small, ubiquitous eukaryotic cell organelles that mediate a wide range of metabolic functions such as photorespiration, fatty acid beta-oxidation and response to biotic and abiotic stress. Recent advances have begun to reveal the unexpectedly large plant peroxisomal proteome to increase our understanding of metabolic pathways in peroxisomes. Large-scale plant genome sequencing will soon allow detailed comparative computational analyses of many different peroxisomal proteomes. The results should be instrumental in defining the functional and metabolic inventory of plant peroxisomes and developing molecular strategies for improvement of food and biofuel production.

We here present the first approach to functional and metabolic characterization of plant peroxisomal proteomes from different phylogenetic clades using bioinformatics and machine-learning methods. Our pipeline involves the prediction of peroxisomal proteins from diverse complete plant genomes followed by a homology search-based identification of clade-specific conserved gene families. By mapping the resulting peroxisomal proteomes to Gene Ontology terms, Pfam domain families and KEGG metabolic pathways we obtain functional and metabolic profiles of different algae, mosses, monocotyledons and dicotyledons. Our computational analyses of metabolic profiles from peroxisomal proteomes and complete genomes reveal significantly enriched peroxisomal pathways that have previously been unknown. Furthermore, we apply machine learning techniques to functional profiles from different clades to identify known and novel discriminative peroxisomal functions and pathways in algae and seed plants. Future work will comprise experimental verification of newly identified proteins and pathways and the extension of our method to other phylogenetic branches and other organelles.

# Different expression of classical Hodgkin lymphoma and primary mediastinal B-cell lymphoma

Denis Dalic[1], Ina Koch[1], Martin-Leo Hansmann[2], Claudia Döring[2]
[1] *Molecular Bioinformatics, Institute of Computer Science, Custer of Excellence Macromolecular Complexes, Johann Wolfgang Goethe-University Frankfurt am Main, Robert-Mayer-Straße 11-15, 60325 Frankfurt am Main, Germany*
[2] *Senckenberg Institute of Pathology, Johann Wolfgang Goethe-University Frankfurt am Main, 60590 Frankfurt am Main, Germany*

Nodular sclerosis is a subtype of classical Hodgkin lymphoma (NScHL). Primary mediastinal B-cell lymphoma (PMBL) is a subtype of diffuse large B-cell non-Hodgkin lymphoma (DLBCL) which is closely related to NScHL. Both types are cancer forms in the lymph system. Although the PMBL is a subtype of DLBCL they are difficult to differentiate. Nevertheless, NScHL and PMBL share many morphologic, immunohistochemical and molecular features.[TDB+12]

We aiming at finding differences or similarities in the metabolism or in the signaling pathways. Therefore, we analyzed gene expression data from five PMBL patients and seven NScHL patients. The first step was to show the different gene expression profile between cancer cells and GC B cells (germinal center B cells) obtained from ten patients undergoing routine tonsillectomy. Using unsupervised hierarchical clustering we found a principal division of PMBL/NScHL and GC B cells. Statistical analysis of the gene expression profiles resulted in 269 significantly over-expressed genes. For these genes we analyze the localization and biological function. Through pathway mapping and enrichment analysis we search now for pathways with a high number of significantly over-expressed genes.

## References

[TDB+12] Enrico Tiacci, Claudia Döring, Verena Brune, Carel JM van Noesel, Wolfram Klapper, Gunhild Mechtersheimer, Brunangelo Falini, Ralf Küppers, and Martin-Leo Hansmann. Analyzing primary Hodgkin and Reed-Sternberg cells to capture the molecular and cellular pathogenesis of classical Hodgkin lymphoma. *Blood*, 120(23):4609–4620, 2012.

# Statistical analysis of Hodgkin lymphoma based on tissue image data

Jennifer Scheidel[1], Tim Schäfer[1], Hendrik Schäfer[1], Jörg Ackermann[1], Claudia Döring[2], Sylvia Hartmann[2], Martin-Leo Hansmann[2], and Ina Koch[1]

[1] *Molecular Bioinformatics, Institute of Computer Science, Cluster of Excellence "Macromolecular Complexes", Johann Wolfgang Goethe-University Frankfurt am Main, Robert-Mayer-Straße 11-15, 60325 Frankfurt am Main, Germany*
[2] *Senckenberg Institute of Pathology, Johann Wolfgang Goethe-University Frankfurt Main, 60590 Frankfurt am Main, Germany*

Nowadays, with cancer diagnosis a large amount of histological images is generated. The systematic analysis of tissue images is an important aspect for studying the development of cancer. Hodgkin lymphoma is a malignant tumor of the lymphatic system, which is of special interest because of its different growing behavior.

For our analysis, we use tissue section images of Hodgkin lymphoma. The histological images are from tissue samples from 58 patients with Hodgkin lymphoma subtype *nodular sclerosis* or *mixed type*, as well as from tissue samples from 24 patients suffering from inflammation of the lymph nodes called *lymphadenitis*. At first, we preprocess and segment the tissue section images in CD30 positive regions as described in [SSS+13]. Secondly, the image analysis detects and classifies the CD30 immunostained cell profiles into different classes according to morphological properties.

We analyzed the class distribution, focusing on the statistical analysis of the neighborhood of a cell profile with a specific morphology. Thus, we detect correlations in the neighborhood relations of classes. Specific classes are favored or avoided in the neighborhood of other classes. For example, in Hodgkin lymphoma the morphological class of small and round cell profiles prefers the neighborhood of themselves and is neutral or even avoids the neighborhood of other classes. Furthermore, we compare the results for the Hodgkin lymphoma subtypes nodular sclerosis and mixed type as well as lymphadenitis and find differences. We observed that the classes of lymphadenitis has less correlations in the neighborhood relations than the classes of Hodgkin lymphoma subtypes.

## References

[SSS+13] Tim Schäfer, Hendrik Schäfer, Alexander Schmitz, Jörg Ackermann, Norbert Dichter, Claudia Döring, Sylvia Hartmann, Martin-Leo Hansmann, and Ina Koch. Image database analysis of Hodgkin lymphoma. *Computational biology and chemistry*, 46:1–7, 2013.

# Analysis of RNA-seq data for identifying flowering time regulators in vernalized and non-vernalized rapeseed

Weinholdt C.[1], Emrani N.[2], Lemnian I.[1], Jedrusik N.[2], Molina C.[2], Jung C.[2], Grosse I.[1]
*(1) Institute of Computer Science , Martin Luther University of Halle-Wittenberg, Germany*
*(2) Plant Breeding Institute, Christian Albrechts University of Kiel, Germany*
claus.weinholdt@informatik.uni-halle.de

Rapeseed is one of the most important oil crops, and the identification of genetic factors for floral induction, which are key regulators in seed production, have a high importance in agriculture. We have extracted RNA from pools of Brassica napus under two environmental conditions, vernalized and non-vernalized, at three stages of early plant development. With the goal of identifying flowering time regulators, we have developed a pipeline for de-novo assembly, contig annotation, and prediction of differentially expressed contigs. For both vernalized and non-vernalized plants, we have found about 500 contigs with strongly varying expression during vegetative to reproductive transition, and we are currently testing these putative transcripts by qPCR.

# Towards an optimal transcriptome assembly of the Naked Mole Rat

Martin Bens, Karol Szafranski, Matthias Platzer
*Leibniz Institute for Age Research-FLI, Jena, Germany*
mbens@fli-leibniz.de

Naked mole rats (NMRs) (*Heterocephalus glaber*) are mouse-sized subterranean rodents with an exceptional long lifespan of >30 years in captivity. During their lifespan they show no age-related decline in fertility and not the typical gradual increase in mortality. Additionally, cancer has never been observed in this species. On the basis of these attributes, investigation of NMRs offers the possibility to discover molecular mechanisms, which lead to a long and healthy lifespan[Aus09, Buf08]. Advances in RNA-seq technology and transcriptome assembly provide a cost-effective way to analyse transcript expression profiles, despite the absence of a fully sequenced genome.

We present a species independent transcriptome assembly pipeline, which utilizes different tools in order to pre-process mRNA-seq reads (sickle, cutadapt[Mar11]), assemble the transcriptome (Trinity[G$^+$11]), screen for contamination (own scripts), reduce redundancy by sequence clustering (CD-HIT[LG06],TGICL[P$^+$03]), identify chimeric transcripts (own scripts) and annotate the assembled transcript contigs (Blast[A$^+$90]). Additionally, scaffolding of transcript contigs based on transcript data of a closely related species is performed and descriptive assembly statistics are reported.

We obtained transcriptome data of 8 different tissues from NMRs, resulting in ∼360 million reads in total. With the aid of our pipeline we find NMR counterparts to 80% of human genes (NCBI Homo Sapiens Annotation Release 104). Scaffolding improved the length of 12% of all annotated genes. NMR counterparts cover 97% of all Gene Ontology terms, 96% of gene families defined by HUGO Gene Nomenclature Committee and 95% of genes annotated in GenAge[dM$^+$09]. A comparison with the current transcript annotation of the NMR genome [K$^+$11] will be presented.

## References

[A+90] SF Altschul et al. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 10 1990.

[Aus09] S N Austad. Comparative biology of aging. *J. Gerontol. A Biol. Sci. Med. Sci.*, 64(2):199–201, February 2009.

[Buf08] Rochelle Buffenstein. Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. *J Comp Physiol B*, 178(4):439–445, May 2008.

[dM+09] J. P. de Magalhaes et al. The Human Ageing Genomic Resources: online databases and tools for biogerontologists. *Aging Cell*, 8(1):65–72, Feb 2009.

[G+11] Manfred G Grabherr et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, 29(7):644–652, July 2011.

[K+11] Eun Bae Kim et al. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, 479(7372):223–227, November 2011.

[LG06] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–9, Jul 2006.

[Mar11] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

[P+03] Geo Pertea et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, 19(5):651–652, 2003.

# Distinction of Type 2 diabetes using PCA, miRNA as features

Shodai Katsukawa and Y-h. Taguchi
*Department of Physics, Chuo University*
sh.ktkw@gmail.com, tag@granular.com

Type II diabetes (T2D) was a critical symptom also related to other diseases including cancers, and has affected health of numerous number of people world wide[1]. Especially, developing thus population holding countries are expected to experience rapid increase of T2D patients because of drastically changing food style caused by economic development. Thus, suppression of T2D is emergently important. A objective of this study is to propose T2D discrimination method by using miRNA as features. In this study, we investigated microRNAs as candidates of drug targets and biomarkers. Because microRNA was regarded to be a critical factors for T2D [2], we propose to use mainly principal component analysis (PCA) for discrimination. A recently proposed principal component analysis based linear discriminant analysis allows us to specify numerous miRNAs differently expressed between type II diabetes patients and healthy control[3]. First of all, 58 miRNAs used for discrimination was selected by PCA. Secondly, we applied PCA to embed patients using selected miRNAs. Finally, they were discriminated using linear discriminant analysis (LDA) using semi-supervised learning. As a result, using this discrimination method, we were able to discriminate T2D, normal control (CTL) and Impaired Fasting Glucose (IFG) patients with the accuracy of 88%. This demonstrates the usefulness of our method. Then, we also researched about contribution to T2D of each miRNAs. miR-144 and 144* should be mostly coincident with progress of diseases. On the other hand, miR-30d should describe difference between T2D and IFG.

## References

[1] Lin, Y. and Sun, Z.: Current views on type 2 diabetes, *J. Endocrinol.*, Vol. 204, No. 1, pp. 1–11 (2010).

[2] Karolina, D. S., Armugam, A., Tavintharan, S., Wong, M. T., Lim, S. C., Sum, C. F. and Jeyaseelan, K.: MicroRNA 144 impairs insulin signaling by inhibiting the expression of insulin receptor substrate 1 in type 2 diabetes mellitus, *PLoS ONE*, Vol. 6, No. 8, p. e22839 (2011).

[3] Taguchi, Y.-h. and Murakami, Y.: Principal Component Analy- sis Based Feature Extraction Approach to Identify Circulating mi- croRNA Biomarker, *PLoS ONE*, Vol. 9 (online), DOI: 10.1371/jour- nal.pone.0066714 (2013).

# Multiple Protein Alignment using Domain Information

Layal Al Ait and Burkhard Morgenstern
*University of Göttingen, Department of Bioinformatics, Goldschmidtstr. 1, 37077 Göttingen, Germany* layal@gobics.de

Most approaches to multiple sequence alignment rely on primary sequences as the only source of information. External sources of information, however, can give valuable hints to possible sequence homologies that may not be obvious from primary-sequence comparison alone. This idea has been implemented in Clustal Ω where *external profile alignments* can be used to improve the quality of the resulting alignments.

Recently, we proposed different approaches to automatically integrate protein domain information into the multiple-alignment procedure [ACM12]. In a first step, we use *HMMER* to search the input sequences against the *PFAM*[FMT+10] database to identify possible protein domains. Segments of the sequences matching to the same *PFAM* domains should then be preferentially aligned. To do so, we use two options. (*a*) With a previously developed *anchoring* option in *DIALIGN*, this program can be *forced* to align matching *PFAM* domains. (*b*) Domain hits and primary-sequence similarity can be integrated into a multiple alignment using a recently developed graph-theoretical optimization algorithm.

Test runs on *BAliBASE* and *SABmark* show that these approaches lead to improved multiple alignments, compared to alignments based on primary-sequence similarity alone. Our tool is available online at

`http://dialign-pfam.gobics.de/SequenceAlignment/`

## References

[ACM12] Layal Al Ait, Eduardo Corel, and Burkhard Morgenstern. Using protein-domain information for multiple sequence alignment. In *Proceedings of the IEEE 12th Int. Conf. on BioInformatics and BioEngineering (BIBE 12)*, pages 163–168, 2012.

[FMT+10] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, et al. The Pfam protein families database. *Nucleic acids research*, 38(suppl 1):D211–D222, 2010.

# Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as drug target candidates

Ryoichi Kinoshita[1] and Mitsuo Iwadate[2] and Hideaki Umeyama[2] and Y-h. Taguchi[1*]
[1] *Department of Physics and* [2] *Department of Biological Science, Chuo University, Tokyo 112-8551, Japan*
*tag@granular.com

**Background** Aberrant DNA methylation is often associated with cancers. Thus, screening genes with cancer associated aberrant DNA methylation is useful method to identify candidates of cancer causing genes. However, aberrant DNA methylation is also genotype dependent. Thus, selection of genes with aberrant DNA methylation dependent upon genotype in cancers is potentially important for tailor-made medicine. The selected genes can be important candidates for drug target.

**Methods** Recently proposed principal component analysis based selection of genes with aberrant DNA methylation[IUIT13] was applied to genotype and DNA methylation patterns in squamous cell carcinoma (SCC) measured with Single Nucleotide Polymorphism (SNP) array.

**Results** SNP frequent in cancers were also highly methylated. Thus, genes with genotype specific DNA methylation will be therapeutic candidates. Selected genes were also previously reported to be related to cancers. Tertiary structures of proteins were also successfully inferred using two profile based protein structure server, FAMS and phyre2. In addition to this, we sought drug candidate compounds using chooseLD among those in DrugBank.

**Conclusions** We have found genes with genotype specific DNA methylation in SCC that can be candidates of drug targets. Many drug candidates compounds were successfully inferred *in silico*.

## References

[IUIT13] S. Ishida, H. Umeyama, M. Iwadate, and Y. H. Taguchi. Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure

Prediction with FAMS for Drug Discovery. *Protein Pept. Lett.*, Jul 2013.

# A longitudinal transcriptome analysis of a fungal aging model indicates that autophagy compensates age-dependent proteasomal impairments

Oliver Philipp[1, 2], Andrea Hamann[2], Jörg Servos[2], Alexandra Werner[2], Heinz D. Osiewacz[2], Ina Koch[1]

[1] *Johann Wolfgang Goethe University, Molecular Bioinformatics, Institute of Computer Science, Faculty of Computer Science and Mathematics & Cluster of Excellence 'Macromolecular Complexes', Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main, Germany*

[2] *Johann Wolfgang Goethe University, Faculty for Biosciences & Cluster of Excellence 'Macromolecular Complexes', Institute of Molecular Biosciences, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany*

o.philipp@bioinformatik.uni-frankfurt.de

Most biological systems are characterized by time-dependent functional impairments known as aging which finally lead to death. The underlying mechanisms are depending on environmental conditions, stochastic processes and the genetic constitution of the individual. In order to systematically study age-dependent changes in gene expression, we performed a genome-wide longitudinal transcriptome analysis (SuperSAGE) of the fungal aging model *Podospora anserina* [Osi13].

In this study, transcript abundance was quantitated for seven days in the lifetime of *P. anserina* – from young to old - which led to more than 10,000 transcript profiles. A fuzzy clustering approach indicates that genes down-regulated during aging are associated with "ribosomes" and the "proteasome", while genes involved in "autophagy" are up-regulated during aging. Subsequently a more stringent analysis was applied where each profile was correlated with time, and only expression profiles were considered which exhibit gradually age tendencies. An enrichment map which provides a network-based intuitive representation method of a gene enrichment result [Mer10] corroborates the findings of the cluster analysis. Hence, we suggest that autophagy, as part of the cellular quality control system, is induced during aging due to impairments of the proteasome system. Furthermore, we found

the transcript levels of genes involved in the cellular energy metabolism, mitochondria, and especially the respiratory chain to fluctuate during aging, but exhibiting strong differences between young and old individuals. Finally, a statistical comparison of our data with a former published transcriptome analysis of a copper depleted mutant [Ser12] confirmed a relationship of the copper metabolism with cellular aging.

## References

[Ser12] Jörg Servos, Andrea Hamann, Carolin Grimm and Heinz D. Osiewacz. A differential genome-wide transcriptome analysis: impact of cellular copper on complex biological processes like aging and development. *PLoS ONE*, 7: 2012.

[Mer10] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili and Gary D. Bader. Enrichment Map: A network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE*, 5: 2010.

[Osi13] Heinz D. Osiewacz, Andrea Hamann and Sandra Zintel. Assessing organismal aging in the filamentous fungus Podospora anserina. *Methods Mol Biol*, 965:439-462, 2013.

# Prediction of Methotrexate Treatment Response in Rheumatoid Arthritis via Affymetrix miRNA Microarray Profiling

Marc Bonin, Stephan Peter, Karsten Mans, Carolin Sohnrey, Gerd-Rüdiger Burmester, Thomas Häupl, Bruno Stuhlmüller

*Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin*
marc.bonin@charite.de

**Introduction**
Methotrexate (MTX) is the standard medication to treat rheumatoid arthritis (RA). Some patients suffer from significant adverse effects and about 30-40% of patients do not respond to MTX. Therefore predictors for treatment response would be useful, which enable the clinician to only treat patients that will benefit from MTX. A novel and promising class of potential biomarkers are miRNAs, which exhibit several advantages.

**Methods and Results**
Whole blood samples of 28 patients (test group) were collected prior to treatment with MTX. A validation patient cohort (n=11) received MTX monotherapy. Extracted total RNA samples were hybridized onto miRNA 2.0 Arrays. miRNA biomarker candidates to predict MTX treatment were determined in the test group and then validated in the validation group to asses the applicability of the miRNA predictor set. Differential expression of human miRNAs among four non-responders versus 14 good responders of the test group was determined by calculating the fold changes and p-values for the miRNA expression signals. In a second step all responders and non-responders from the test and the validation group were used to predict MTX response with a nearest prototype classification model. A leave-one-out cross-validation was performed instead of splitting the data into training and test group. The test groups consisted of active early RA patients with a DAS28 >5.1 and a disease duration <1 year. The validation group showed a disease duration <2 years. With reference to the MTX response prediction the leave-one-out cross-validation showed sensitivities up to 65 % and specificities up to 95 %.

**Conclusion**
Prediction of response to MTX therapy using microarray analyses allows reducing costs, preventing side effects and is an opportunity for effective ‘individualized medicine’. Next to mRNA expression analysis the investigation of miRNAs as posttranscriptional regulators could prove helpful to better understand physiological and pathological processes, to define miRNA biomarkers and predict response to medication.

# Chronic Inflammation is associated with cancer-related methylation changes

Sebastian Bender[1], Monther Abu-Remaileh[2], Günter Raddatz[1], Yehudit Bergman[2], Frank Lyko[1]

[1]*Department of Epigenetics, DKFZ Heidelberg*

[2]*Department of Developmental Biology and Cancer Research, The Hebrew University Jerusalem*

sebastian.bender@dkfz-heidelberg.de

Cancer is a disease that is mainly caused by the misregulation of genes (sometimes copy number variation or single nucleotide polymorphisms), leading to uncontrolled growth/proliferation, loss of cell function, etc. Epigenetics studies changes in gene expression or phenotype caused by modifications other than the pure underlying DNA sequence, making it an important part of cancer research. Some cancer studies implicated inflammation as the cause or a reinforcing factor of cancer formation. The focus in this field is mainly DNA methylation, but also histone modification, transcription factor binding or RNA interference.

In order to investigate the role of inflammation and DNA methylation in cancer formation, we performed a Whole Genome Bisulfite Sequencing (WGBS) for a common chemically induced mouse colon cancer model[WNWN07], comparing wildtype to samples treated with DSS (causing inflammation) and AOM+DSS (causing cancer and inflammation). For the analysis on a grand scale, we chose to search for methylation differences in major structural features of mammalian methylomes, partially methylated domains (PMDs) and DNA methylation valleys (DMVs). In contrast to other cancer studies, the PMDs were not very pronounced, but the inflamed and cancerous methylation patterns were remarkably similar. The DMV analysis showed some interesting correlations between gene expression and differential DMV methylation. While many differentially methylated DMVs were in concordance in cancer and inflammation tissues, we also found some important differences associated with genes shown to be colon cancer related.

## References

[WNWN07] Wirtz, S., Neufert, C., Weigmann, B. & Neurath, M. F. Chemically induced mouse models of intestinal inflammation. *Nature protocols* **2**, 541–6 (2007).

# Identify cell line specific microRNA TSS based on H3K4m3 data

Xu Hua[1,2], Jie Li[1], Jin Wang[2], Edgar Wingender[1]
[1]*Institute of Bioinformatics, University Medical Center Goettingen*
[2]*School of Life Science, Nanjing University*
xu.hua@bioinf.med.uni-goettingen.de

MicroRNAs (miRNA) are crucial small non-coding RNAs which play important roles in various biological disease processes by repressing or inhibiting transcription of mRNAs. Nowadays, it is still problematic to identify miRNA TSS (transcriptional start site), because the 5' cap of primary miRNA transcript is rapidly processed after the transcription, which consequently renders it unsuitable to locate miRNA TSSs precisely with the cap-dependent experiments.

Here we developed a strategy for the identification of miRNA TSSs in 54 cell lines based on H3K4m3 data [1] which gives the indication of a TSS locus. For every nearest H3K4m3 segment to a pre-miRNA in a certain cell line, we considered the features of TSS in sequence and conservation, employing Eponine (a widely-used TSS-prediction tool for protein-coding gene) [2] and a high-conservation criteria, to precisely locate the miRNA TSS. Comparing to previous identifications of 6 miRNA TSSs, which were verified by experiment, our method gives the TSSs that are closest to the experimentally verified ones. We also plotted the conservation distribution around our predicted TSSs, which confirmed the functional importance of the predicted regions. In addition, the generally low distance from our predicted TSS to the start point of the pre-miRNA also suggests the better performance than previous methods.

## References

1. The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. Nature. 489, 57–74, 2012
2. Down, T.A. and Hubbard, T.J., Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. Genome Res. 12:458-461, 2002.

# Protein Folding and Structure through Synchronization

Leandro P. Nadaletti, Beatriz S. L. P. de Lima, and Solange Guimarães
*Graduate School in Engineering, Federal University of Rio de Janeiro (UFRJ)*
lnadaletti@coc.ufrj.br

Different models such as diffusion-collision and nucleation-condensation have been used to unravel how secondary and tertiary structures form during protein folding. Although such models enhance our understanding of the problem, we have still not identified a simple mechanism based on physical principles that generally describes folding kinetics and thermodynamics. The synchronization of the dynamics of amino acids is a mechanism that adequately explains how a protein rapidly adopts a specific three-dimensional native state structure. Synchronization is the ability for self-organization, wherein two or more self-sustaining dynamic systems adjust their rhythms to adopt coordinated behaviour through a low-intensity mutual influence [Pik01]. Such non-linear phenomena are widespread in nature and are applied in different fields of study, such as modelling cardiac cell behaviour in biology, and superconductors in physics, through Josephson junction arrays. We established a parallel between the folding process and dynamics for a network of coupled oscillators described by the Kuramoto model [Str00] to illustrate the inherent concepts of the new proposed model. Amino acid coupling explains the mean-field character of the force that propels an amino acid sequence into a structure through self-organization. Therefore, synchronization can be seen as a unifying principle for the different protein folding models, since different methods for reaching global synchronization reproduce the folding dynamics characteristics of proteins.

## References

[Pik01] A. Pikovsky et al., Synchronization: A Universal Concept In Nonlinear Sciences, UK:Cambridge University Press, 2001.
[Str00] S. H. Strogatz, "From kuramoto to Crawford", Physica D: Nonlinear Phenomena, vol. 43, no. 1, str. 1-20, 2000.

# Network of Silence

Stephan Flemming[1], Simon Bohleber[1], Thomas Häupl[2], Stefan Günther[1]
[1] *Institute of Pharmaceutical Sciences, Pharmaceutical Bioinformatics, Albert-Ludwigs-Universität Freiburg*
[2] *Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin*
*stephan.flemming@pharmazie.uni-freiburg.de*

DNA methylation is the covalent binding of a methylgroup to cytosins within a cytosin-guanin dinculeotide (CpG site). The approximately 28 million CpG sites in the human genome are not equally distributed and occur in clusters of high CpG density, called CpG islands (CGI). CGI can be found in approximately 60% of all human gene promoters and are also located in sections surrounding the transcription start site (5'UTR), gene body and sections that follow the translation terminal codon (3'UTR). [DDC+11] CpG dinucleotides can be methylated, unmethylated and hemimethylated. Changes in methylation levels of CpG dinucleotides correlate with transcriptional repression and gene silencing, although not all sites have the same impact on gene expression. [FZ09] The aim of this study is the identification and characterisation of CpG sites which are more predictive for changes in gene expression than others. The Illumina HumanMethylation450 Beadchip platform provides a genome-wide coverage of more than 450,000 CpGs for more than 23,000 genes. Based on the analysis of $\approx$ 5,300 samples categorized in $\approx$ 120 series derived with the Illumina HumanMethylation450 platform [DDC+11], network analysis approaches and machine learning techniques are applied to identify groups of CpG sites with similar methylation patterns and functional properties.

## References

[DDC+11] Sarah Dedeurwaerder, Matthieu Defrance, Emilie Calonne, Hélène Denis, Christos Sotiriou, and François Fuks. Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, 3(6):771–84, December 2011.

[FZ09] Shicai Fan and Xuegong Zhang. CpG island methylation pattern in different human tissues and its correlation with gene expression. *Biochemical and biophysical research communications*, 383(4):421–5, June 2009.

# Prediciting Alzheimer Disease using miRNA signatures

Jerzy Dyczkowski, Pooja Rao, Angela Dettmar, Anja Schneider, Andre Fischer, Stefan Bonn

*German Center for Neurodegenerative Diseases (DZNE)*
jerzy.dyczkowski@dzne.de

Small non-coding RNAs (miRNAs) are important regulators of many cellular processes including cancer progression and various neurological diseases. Recent evidence suggests that miRNA signatures could potentially be used as prognostic biomarkers for diseases of the brain, as has been shown for example for schizophrenia and brain tumors. Here, we propose the potential use of miRNA signatures derived from blood and cerebrospinal fluid (CSF) for the diagnosis of Alzheimer Disease, the most prevalent neurodegenerative disease in humans. We analysed blood- and CSF-derived miRNA profiles from Alzheimer Disease and control patient material using next generation sequencing. MiRNA counts were generated using a custom analysis pipeline that combines blast search on miRNA databases and whole genome alignment. Normalized miRNA counts were stratified for confounding factors using linear models and then subjected to machine learning for model building. The stratification for tissue-type, patient age, and standard operating procedures/source of the sample was crucial to obtain reliable prognostic disease signatures. Using an optimal set of 18 miRNAs, Alzheimer Disease could be predicted with 82% accuracy in CSF samples. In summary, we identified a predictive miRNA signature for Alzheimer Disease and give guidelines for the proper derivation of disease signatures from miRNA-seq data.

# Genotype-phenotype correlation of continuous characters while considering phylogeny

Amol Kolte and Farhat Habib
*Center of Excellence in Epigenetics, Indian Institute of Science Education and Research, Pune - 411008*
fhabib@iiserpune.ac.in

Despite rapid growth in the availability of sequence data, due to advent of next generation sequencing technologies, and phenotype data the mapping between genotypes and phenotypes is poorly known especially when it comes to complex traits. The significance of association between two characters in a group of organisms can be influenced by the interrelationships between them as described by their phylogeny [Fel85]. Previously, we described algorithms that could produce genotype-phenotype correlations while considering the phylogeny but were limited to binary or discrete characters [HJBJ07]. We describe a method, `Gephcort`, to find correlations between genotypes and a continuous phenotype taking phylogeny into consideration.

The method relies on finding the optimal states of the genotypes and phenotype at the internal nodes of the given phylogeny and then finding whether changes in the genotype and phenotype are co-occuring over the branches of the tree. Randomization testing is used to assess the significance of the correlation between the genotype and the phenotype. As a case study, we correlate High Density Lipoprotein Cholesterol levels in inbred mice with their genotype. Comparison of our results with literature surveys of previous *in silico* and experimental studies for this trait shows that our method compares favourably and can be used to infer genotype-phenotype associations rapidly compared to wet-lab methods for inferring associations. `Gephcort` has been implemented in the Python programming language and is available under GPL from https://github.com/farhat.

## References

[Fel85] J. Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125(1):1–15, 1985.

[HJBJ07] Farhat Habib, Andrew D. Johnson, Ralf Bundschuh, and Daniel Janies. Large scale genotype-phenotype correlation analysis based on

phylogenetic trees. *Bioinformatics*, 23(7):785–788, 2007.

# Finding Functional Interactions of Proteins and Small Molecules in Sentences of PubMed Abstracts

Kersten Döring, Michael Becer, and Stefan Günther
*Institute of Pharmaceutical Sciences, Albert-Ludwigs-Universität Freiburg*
kersten.doering@pharmazie.uni-freiburg.de

Current knowledge about protein-chemical interactions is essential for the development of new drugs. This information is provided by sources of experimental information, manually curated drug targets, and pathway databases. Especially new information can be extracted from literature by using co-occurrence text mining and Natural Language Processing [K+12]. There are many different technical approaches for automated protein-protein interaction extraction from unstructured text collections related to machine learning and feature selection [S+11]. We have applied these methods within the field of protein-chemical interactions. 20,000 PubMed abstracts have been searched for sentences including at least one protein and one small molecule by using the web service *Protein-Literature Investigation for Interacting Compounds* (prolific) [SG+12]. Phrases containing a *relationship word* enclosed by two biomolecules have been analysed and classified as *interaction* or *no interaction* instances. The definition of an interaction includes direct binding of two biomolecules as well as indirect relationships such as increase of protein expression. We show the results of a classifier filtering out co-occurrences that match this definition.

## References

[K+12] Michael Kuhn et al. STITCH 3: Zooming in on Protein-Chemical Interactions. *Nucleic Acids Research*, 40(D1):D876–80, 2012.

[S+11] Min Song et al. Combining Active Learning and Semi-Supervised Learning Techniques to Extract Protein Interaction Sentences. *BMC Bioinformatics*, 12(Suppl 12):S4, 2011.

[SG+12] Christian Senger, Björn A. Grüning, et al. Mining and Evaluation of Molecular Relationships in Literature. *Bioinformatics*, 28(5):709–14, 2012.

# A parametric analyse of the asymmetric Wagner parsimony

Gilles Didier
*Institut de Mathématiques de Luminy, Aix-Marseille Université*
*Campus de Luminy, Case 907, 13288 MARSEILLE Cedex 9*
gilles.didier@univ-amu.fr

The Wagner parsimony, a.k.a. linear parsimony, yields to reconstruct the ancestral states of a quantitative character (size, volume etc.) from its values on extant taxa, by minimizing the sum of absolute differences between the states of all the nodes and their direct descendants. This approach implicitly assumes that the character evolution is neutral since an increase lead to a same cost as a decrease of the same amount. If one wants to take into account an evolutionary trend, a natural way is to define an asymmetric cost of evolution $\Delta$ between an ancestor and its direct descendant [Csu08]:

$$\Delta\{y \to x\} = \begin{cases} \gamma(x-y) & \text{if } y < x \\ \lambda(y-x) & \text{if } y > x \end{cases}$$

where $y$ and $x$ are the ancestor and descendant states, respectively.

Like in [Did11], we study the influence of the parameters $\gamma$ and $\lambda$ on the states reconstructed by the method. First, we show that whatever $\gamma$ and $\lambda$, there always exists a most parsimonious reconstruction in which all the states are taken from the set of known states (that of extant taxa). Next, we provide an algorithm giving, from a phylogenetic tree and the states of extant taxa and for each ancestral node, the states reconstructed with regard to the ratio $\gamma/\lambda$. Finally, the approach is applied on two biological datasets.

## References

[Csu08] Miklós Csurös. Ancestral Reconstruction by Asymmetric Wagner Parsimony over Continuous Characters and Squared Parsimony over Distributions. In Craig Nelson and Stéphane Vialette, editors, *Comparative Genomics*, volume 5267 of *Lecture Notes in Computer Science*, pages 72–86. Springer Berlin / Heidelberg, 2008.

[Did11] G. Didier. Parametric Maximum Parsimonious Reconstruction on Trees. *Bulletin of Mathematical Biology*, 73:1477–1502, 2011.

# Comparison of protein topology graphs using graphlet-based methods

Tatiana Bakirova, 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*

*bakirova.tatiana@gmail.com tim.schaefer, ina.koch@bioinformatik.uni-frankfurt.de*

With the rapidly growing amount of protein structures available in public databases like the RCSB Protein Data Bank, there is a strong need for developing fast and accurate comparison methods for proteins. Detecting functional similarity, evolutionary relationships and/or structural motifs at different description levels of proteins is of major interest for many applications in biology and molecular medicine. Although there already exist different methods for protein structure alignment, the search of protein structure databases is still time consuming and similarity between proteins can be defined in many ways. Common methods used to compare proteins include sequence-based tools like BLAST and databases like CATH and SCOP, which also consider 3Ddata like the topological arrangements of secondary structure elements. The PTGL is a protein topology database [SMK12] which uses a graph-based model to describe protein structure on the secondary structure level. We present an approach of comparing proteins which allows to retrieve similar proteins from this database. Our method is based on the graphlet counting algorithms for unlabeled graphs [She+09] which were extended by counting several types of labeled graphlets and by the similarity model based on these counts.

**References**

[She+09] Shervashidze, N., Petri, T., Mehlhorn, K., Borgwardt, K. M., & Viswanathan, S. (2009). Efficient graphlet kernels for large graph comparison. International Conference on Artificial Intelligence and Statistics (pp. 488-495).

[SMK12] Schäfer, T., May, P., & Koch, I. (2012). Computation and Visualization of Protein Topology Graphs Including Ligand Information. In GCB (pp. 108-118).

# Predicting targets of synergistic microRNA regulation

Ulf Schmitz[1*], Shailendra Gupta[1], Xin Lai[1,2], Julio Vera[2] and Olaf Wolkenhauer[1]

*1 Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany*

*2 Department of Dermatology, University of Erlangen-Nurnberg, Erlangen, Germany*

[*]ulf.schmitz@uni-rostock.de

MicroRNAs (miRNAs) regulate gene expression in mammals at the post-transcriptional level and in most of cell-biological processes. Saetrom and colleagues characterized circumstances under which miRNA pairs can repress the translation of a mutual target mRNA in a cooperative manner [1]. In our own work we have shown that higher quantities of miRNAs, as well as the phenomenon of synergistic target regulation, can realize efficient noise buffering for external stimuli, triggering target expression [2].

Here, we present a computational workflow to identify new targets of synergistic miRNA regulation and to characterize triplexes composed of two cooperating miRNAs and a target mRNA. This workflow includes six steps that iteratively increase the level of detail and confidence in the functionality of the predicted synergistic target regulation: (i) identification of miRNA binding sites in the 3′ UTR of target genes; (ii) identification of putatively cooperating miRNAs with binding sites in close proximity; (iii) prediction and analysis of the local secondary structure of putative RNA-triplexes; (iv) determination of the triplex thermodynamic profile; (v) determination of binding affinities among the involved molecules; and (vi) simulation of target steady-state values under influence of cooperating miRNAs.

We have analysed the ‘whole human genome’ and identified 17,259 putative targets of synergistic repression by a total of 249 miRNAs.

Our analysis shows that the local structure formed by two miRNAs and their mutual target determines the strength of synergistic miRNA regulation. The proposed workflow provides a comprehensive way to predict targets of

cooperative miRNA regulation and the target repression efficiency with high confidence.

## References

1. Saetrom P, Heale BSE, Snøve O, Aagaard L, Alluin J, Rossi JJ: Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res 2007, 35:2333–2342.
2. Lai X, Schmitz U, Gupta SK, Bhattacharya A, Kunz M, Wolkenhauer O, Vera J: Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs. Nucleic Acids Res 2012, 40 (18): 8818-8834.

# GEDEVO: An Evolutionary Graph Edit Distance Algorithm for Biological Network Alignment

Rashid Ibragimov[1], Maximilian Malek[1], Jiong Guo[2], and Jan Baumbach[1,3]
[1] *Max-Planck-Institut für Informatik,*
[2] *Universität des Saarlandes,* [3] *University of Southern Denmark*
{ribragim,mmalek}@mpi-inf.mpg.de
jguo@mmci.uni-saarland.de, jan.baumbach@imada.sdu.dk

Introduction: With the so-called OMICS technology the scientific community has generated huge amounts of data that allow us to reconstruct the interplay of all kinds of biological entities. The emerging interaction networks are usually modeled as graphs with thousands of nodes and tens of thousands of edges between them. In addition to sequence alignment, the comparison of biological networks has proven great potential to infer the biological function of proteins and genes. However, the corresponding network alignment problem is computationally hard and theoretically intractable for real world instances.

Results: We therefore developed GEDEVO, a novel tool for efficient graph comparison dedicated to real-world size biological networks. Underlying our approach is the so-called Graph Edit Distance (GED) model, where one graph is to be transferred into another one, with a minimal number of (or more general: minimal costs for) edge insertions and deletions. We present a novel evolutionary algorithm aiming to minimize the GED, and we compare our implementation against state of the art tools: SPINAL, GHOST, C-GRAAL, and MI-GRAAL. On a set of protein-protein interaction networks from different organisms we demonstrate that GEDEVO outperforms the current methods. It thus refines the previously suggested alignments based on topological information only.

Conclusion: With GEDEVO, we account for the constantly exploding number and size of available biological networks. The software as well as all used data are publicly available at `http://gedevo.mpi-inf.mpg.de`. See also [IMGB13] for more details.

## References

[IMGB13] Rashid Ibragimov, Maximilain Malek, Jiong Guo, and Jan Baumbach. GEDEVO: An Evolutionary Graph Edit Distance Algorithm for Biological Network Alignment. In *GCB*, 2013.

# Nonlinear Methods of DNA Coding Regions Identification

Vyacheslav A. Tykhonov and Nataliia V. Kudriavtseva
*Department of Radio-electronic Systems, Kharkiv National University of Radioelectronics, Kharkiv, Ukraine*
E-mail: slavatihonoff@mail.ru, nataliia.kudriavtseva@gmail.com

Spectral analysis is traditionally used to identify protein coding regions [1, 2]. A new method of DNA parametric power spectrum density (PSD) estimations for identification of protein coding regions has been proposed in this work for the first time. For analysis of non–Gaussian processes Fourier third, fourth, fifth and sixth orders statistics of a DNA sequence have been used after replacement of some codons to the synonyms. Autoregressive (AR) filter was used for smoothing of the DNA sequence for the first time. Smoothening AR filter allows to decrease fluctuation of the spectrograms and to intensify the spectrograms on $N/3$ frequencies. To improve the observability of all spikes and, therefore, of coding DNA regions, limited by the level spectrum have been received for the first time. When spectrum's order is rising, we can see an equal rise of all spectrogram's spikes with respect to the noise level. The signal-to-noise ratio (SNR) for the proposed method of analysis and the determination accuracy of coding and non-coding DNA regions have been estimated. SNR depends on the clip's level. Dependencies of SNR on the spectrum's order using clips 0.4, 0.6, 0.7 have been received. The analysis of the results shows, the estimation accuracy, the noise level, spectra' shapes are significantly high than in the best work known in this field of research [1]. Maximal value of SNR in this work is equal to 14, in other papers with the same research it is equal to 1.35 [2]. The parameter p that shows determination accuracy of coding and non-coding DNA regions was received equal to more than 0.95. In other papers they received this parameter equal to near 0.91 [3].

## References

[1] Dimitris Anastassion, "Genomic signal processing", In: IEEE Signal processing magazine, 2001, pp.8-20.
[2] A.Khare, A. Nigam, M.Saxena. Identification of DNA Sequences By Signal Processing Tools in Protein-Coding Regions. Search & Research. Vol-II No.2 (44-49):2011. Pp. 44-49.
[3] T.S. Gunavan, E. Ambkairajah, J. Epps. A Signal Boosting Technque for Gene Prediction. ICICS 2007. IEEE.

# Basic topological features for metabolic pathway models

Jens Einloft, Jörg Ackermann and Ina Koch
*Goethe University Frankfurt, Institute for Computer Science, Molecular Bioinformatics*
Einloft@bioinformatik.uni-frankfurt.de

The development and analysis of metabolic networks an / or signal transduction networks is an important topic in systems biology. Graph theory, which networks are based on, is a subject of intensive research since the 18th century, and concerns, besides other topics, with the meaning of topological properties. In the raising importance of the analysis of biological networks in the last decades, researchers are aiming to find correlations between topological properties and biological interpretation. A first investigation of biological networks, concerning their node degree distribution, was done by [JTA+00], suggesting scale free network structure in biological networks. [RSM+02] analyzed properties like the cluster coefficient of metabolic networks. Both studies are based on the same dataset of 43 metabolic networks. The common hypergraph representation of the networks limits the investigation to the properties of the metabolites, without considering the impact of the reactions on these properties. In contrast, bipartite graphs allow to explore the effect of the reactions on the topological properties. We choose the Petri net models that characterize the topological properties of both, reactions and metabolites. We focus on three topological properties: node degree, cluster coefficient, and the shortest path length. Our study is based on 1846 different whole-genome metabolic networks from the path2models database. We found that in this graph representation the node degree of metabolites follow, in contrast to reactions, a scale free distribution. Also, we show the big influence of the secondary metabolites on these distributions. The small world property applies for both, but only because of the secondary metabolites.

## 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.

# Prediction of protein interaction types based on sequence and network features

Florian Goebels[1] and Dmitrij Frishman[1,2]

[1]*Department of Genome Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Germany*

[2]*HMGU German Research Center for Environmental Health, Institute for Bioinformatics and Systems Biology/MIPS, Germany*

d.frishman@wzw.tum.de

Protein interactions mediate a wide spectrum of functions in various cellular contexts. Functional versatility of protein complexes is due to a broad range of structural adaptations that determine their binding affinity, the number of interaction sites, and the lifetime. In terms of stability it has become customary to distinguish between obligate and non-obligate interactions dependent on whether or not the protomers can exist independently. In terms of spatio-temporal control protein interactions can be either simultaneously possible (SP) or mutually exclusive (ME). In the former case a network hub interacts with several proteins at the same time, offering each of them a separate interface, while in the latter case the hub interacts with its partners one at a time via the same binding site. So far different types of interactions were distinguished based on the properties of the corresponding binding interfaces derived from known three-dimensional structures of protein complexes. Here we present PiType, an accurate 3D structure-independent computational method for classifying protein interactions into SP and ME as well as into obligate and non-obligate. Our classifier exploits features of the binding partners predicted from amino acid sequence, their functional similarity, and network topology. We find that the constituents of non-obligate complexes possess a higher degree of structural disorder, more short linear motifs, and lower functional similarity compared to obligate interaction partners while SP and ME interactions are characterized by significant differences in network topology. Each interaction type is associated with a distinct set of biological functions. Moreover, interactions within multi-protein complexes tend to be enriched in one type of interactions.

# Simultaneous Gene Prediction in Related Species

Stefanie König, Lizzy Gerischer and Mario Stanke
*Institute of Mathematics and Computer Science,*
*University of Greifswald*
{stefanie.koenig, lizzy.gerischer, mario.stanke}@uni-greifswald.de

More and more frequently, genome sequencing projects aim to sequence genomes of several closely related species, rather than single genomes. For the identification of protein coding genes in such genomes, it would be convenient to have computational methods that are able to exploit the information given by the evolutionary relationship of the species. Especially with regard to computational complexity this task is very challenging. Common approaches to *de novo* comparative gene finding use a multiple alignment of sequences, e.g. to consider sequence conservation. However, they restrict the prediction and gene structure model to a single reference genome.

We extend the gene finder AUGUSTUS [SDBH08] by a novel comparative approach that simultaneously identifies genes in genomes of multiple species. This means that the prediction of a gene structure in a single species depends on the existence of corresponding gene structures in the other species. Note that our approach does not assume one gene structure common to all of the species. Gene structures across different species may differ, e.g. in the number of exons or positions of splice sites. They may even be completely absent in some of the species.

In graph-theoretic terms, we seek to maximize a score in a graph $G$, that is a sum of partial scores of the nodes (= exons) in $G$. Although this maximization problem is generally NP-hard, our results suggest that in the majority of cases an exact solution can still be found using an approximative, dual decomposition approach.

## References

[SDBH08] M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*, 24(5):637–644, 2008.

# Multiple Protein Alignments – Structure Versus Sequence-Based

Iryna Bondarenko and Andrew E. Torda
*Centre for Bioinformatics, Hamburg University*
bondarenko@zbh.uni-hamburg.de

There are many different methods to make multiple sequence alignments for the phylogenetic analysis of proteins. Since the choice of the method crucially impacts the downstream analysis one should be very careful and aware of the errors they produce. Using the recently developed mathematical metrics[1], we compare different sequence- and structure-based methods. The methods produce completely different alignments, even in cases with >80% sequence identity. In the regime of lower sequence similarity there are two clear outliers. One of them is the pure structure-based Salami[2]. The other is the adaptive statistical approach *Bali-Phy*[3]. Despite the tradition of sequence-based methods, there is strong evidence that they are full of weaknesses and classifications relying on them may be misleading. We will give examples where the use of a standard classification has led to a wrong evolutionary interpretation.

## References

1. Blackburne, B.P & Whelan, S Measuring the distance between multiple sequence alignments. Bioinformatics 28, 495-502, 2012
2. Margraf, T., Schenk, G., & Torda, A.E. The salami protein structure server. Nucleic Acids Res. 37, 480-484, 2009.
3. Redelings, B.D. & Suchard, M.A. Joint bayesian estimation of alignment and phylogeny. Syst. Biol. 54, 401-418, 2005.
4. Abroi, A., Gough, J. Are viruses a source of new protein folds for organisms? – Virosphere structure space and evolution. Bioessays, 2011.

# RNA sequence design and experimental verification

Marco C. Matthies, Kristina Gorkotte-Szameit, Stefan Bienert, Cindy Meyer, Ulrich Hahn, Andrew E. Torda
*Centre for Bioinformatics, University of Hamburg*
*Department of Chemistry, University of Hamburg*
*SIB / Biozentrum, University of Basel*
*Howard Hughes Medical Institute, Rockefeller University*
{matthies,torda}@zbh.uni-hamburg.de

We have developed an RNA sequence-design method based on Newtonian dynamics in sequence space [MBT12]. This was used to design sequences for three- and four-way junction motifs. The secondary structures have now been experimentally tested.

In the design method, one fixes the desired structure and puts an (ACGU) vector at each position. One uses the nearest-neighbour model for energies of secondary structures, adds an entertaining term for negative design and calculates forces in this sequence space. This allows one to use simulated annealing with thermostatted Newtonian dynamics to find well-suited sequences.

Several of the sequences have been synthesised and checked using chemical structure probing with SHAPE (Selective 2'-hydroxyl acylation analysed by primer extension). We have also compared experimental and predicted melting temperatures.

## References

[MBT12] Marco C. Matthies, Stefan Bienert, and Andrew E. Torda. Dynamics in Sequence Space for RNA Secondary Structure Design. *J. Chem. Theory Comput.*, 8:3663–3670, 2012.

# Automated Peak Extraction for MCC/IMS Measurements of Exhaled Breath

Marianna D'Addario, Dominik Kopczynski, Jörg Ingo Baumbach and Sven Rahmann
*Computer Science 11, TU Dortmund*
Marianna.Daddario@tu-dortmund.de

An ion mobility (IM) spectrometer coupled with a multi-capillary column (MCC) measures volatile organic compounds (VOCs) in the air or in exhaled breath. The MCC/IMS technology has several biotechnological and medical applications for breath analysis, such as lung cancer diagnosis [WLM+10], COPD diagnosis [KHS+11], biomarker discovery and disease classification. For these scopes the raw measurement must be reduced to a set of peaks. Peaks represent VOCs (known or unknown) and are described at least by their coordinates (retention time in the MCC and reduced inverse ion mobility in the IMS) and their signal intensity. Finding peaks within an MCC/IMS measurement is a fundamental step and is referred to as peak extraction. Current state-of-the-art peak extraction methods require human interaction, such as pin-pointing peaks, assisted by a visualization of the data matrix.

Because of the increasing number of available datasets, the need emerges for automated peak extraction in MCC/IMS measurements. Within a high-throughput context in breath gas analysis, we introduce an automated modular peak extraction framework consisting of four sequential steps: Preprocessing, Peak Candidate Detection, Peak Picking and Peak Modeling. Each step has well defined input and output formats and is executed by exchangeable modules.

**Acknoledgments** MDA, DK, JIBB, SR 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

[KHS+11] R. Koczulla, A. Hattesohl, S. Schmid, B. Bödeker, S. Maddula, and J. I. Baumbach. MCC/IMS as potential noninvasive technique in the diagnosis of patients with COPD with and without alpha 1-antitrypsin deficiency. *International Journal for Ion Mobility Spectrometry*, 14(4):177–185, 2011.

[WLM+10] M. Westhoff, P. Litterst, S. Maddula, B. Bödeker, S. Rahmann, A. N. Davies, and J. I. Baumbach. Differentiation of chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control group by breath analysis using ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, 13(3-4):131–139, 2010.

# Dinucleotide distance histograms for fast detection of rRNA in metatranscriptomic sequences

Heiner Klingenberg, Robin Martinjak, Frank Oliver Glöckner, Rolf Daniel, Thomas Lingner and Peter Meinicke
*Department of Bioinformatics, University of Göttingen*
peter@gobics.de

With the rise of metatranscriptomics the study of gene expression in microbial communities has become a rapidly growing field of research. However, new challenges are met when analysing environmental RNA-Seq data and new efficient bioinformatics tools are necessary to cope with the bulk of sequence reads. A first step in the analysis of metatranscriptomic sequencing reads is the separation of rRNA and mRNA fragments to restrict functional analysis to protein coding sequences. Different methods for rRNA filtering provide a variable trade-off between speed and accuracy for a particular dataset. We introduce a machine learning approach for the detection of rRNA in metatranscriptomic sequencing reads based on support vector machines in combination with dinucleotide distance histograms for feature representation. Results show that our SVM based method is at least one order of magnitude faster than any existing tool with only a slight decrease in detection performance when compared to state-of-the-art alignment-based methods.

# A Memory Efficient Data Structure for Pattern Matching in DNA 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

Since backward search was introduced by Ferragina and Manzini, it became a standard index-based linear-time low-memory exact pattern search technique [FM00], especially for read mapping applications in the context of next generation sequencing data analysis. When preparing a standard occurence table for backward search using the complete genome and its reverse complement, the table would require about 50GB memory. Read mappers like BWA [LD09] store only every $k-$th entry of the occurence table and compute the missing entries during execution. This method requires to store also the original genome. We propose a new data structure to store the complete information of the occurence table (without storing the genome and its reverse complement additionally) in less than 4GB without losing computation time. Since all values are present, there is no need to restore missing values.
Because all utilized operations are bitwise operations, our approach has the capability to be efficiently implemented in GPGPUs or FPGAs.

**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.

[LD09] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

# Mixture models for the estimation of metagenomic abundances

Kathrin P. Aßhauer, Heiner Klingenberg, Thomas Lingner, and Peter Meinicke

*Department of Bioinformatics, Institute for Microbiology and Genetics, University of Göttingen, kathrin@gobics.de*

Metagenomics has become a standard approach to analyze microbial communities from environmental and clinical samples. The advances of next-generation sequencing technologies allow researchers to investigate the diversity even of complex microbial communities. However, this develop- ment demands new bioinformatics tools which can efficiently deal with metagenomic data sets on a large-scale. We developed Taxy-Pro [KALM13] and a Mixture of Pathways (MoP) model for the estimation of taxonomic and metabolic abundances, repectively. Both methods provide a solid statistical basis and at the same time, a fast computation of taxonomic and metabolic profiles. Taxy-Pro implements a mixture model based on protein domain frequen- cies inferring the taxonomic composition over the whole range of biolog- ical entities and is freely available at http://www.gobics.de/TaxyPro/ as a Matlab/Octave toolbox or through the CoMet web server [LASM11] (http://comet.gobics.de/). The MoP model extends the taxonomic mixture model to a statisti- cally adequate modeling of the metabolic potential of metagenomes. To overcome computationally intense homology searches, we implemented a shortcut to estimate the metabolic profile of a metagenome. Here, we link the taxonomic profile of the metagenome to a set of pre-computed metabolic reference profiles. The combination of the taxonomic abun- dance estimates and the metabolic reference profiles achieves an unrivaled speed of the metabolic profiling approach. Our results on a large-scale analysis of data from the Human Microbiome Project (HMP) show the utility of our method for fast model-based estimation of pathway abun- dances. Further, the results indicate that the pathway abundances pro- vide a good summary of the functional capacity of a microbial commu- nity, well-suitable for the identification of relevant metabolic differences between distinct body sites/microbial communities.

**References**

[KALM13] Heiner Klingenberg, Kathrin Petra Aßhauer, Thomas Lingner, and Peter Meinicke. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. Bioinformatics, 2013.

[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(suppl 2):W518–W523, 2011.

# Modelling NF-κB signal transduction using Petri nets

[1]Leonie Amstein, [1]Nadine Schöne, [2]Simone Fulda, and [1]Ina Koch
[1]*Molecular Bioinformatics Group, Cluster of Excellence "Macromolecular Complexes", Goethe-University Frankfurt am Main*
[2]*Institute for Experimental Cancer Research in Pediatrics, Cluster of Excellence "Macromolecular Complexes", Goethe-University Hospital Frankfurt am Main*
L.Amstein@bioinformatik.uni-frankfurt.de

Tumor necrosis factor receptor 1 (TNFR1) mediates an important pathway in immune response regulation. Binding of tumor necrosis factor-α can either trigger death or survival of the cell. If transcription factor nuclear factor-κB (NF-κB) is activated, the expression of survival genes is enhanced. A dysregulation of signal transduction may result in inflammatory diseases or cancer [Wal11]. Accordingly, this system requires a strictly controlled regulatory network to conduct the cell response following TNFR1 stimulation. To elucidate the dynamics and regulations, we develop a Petri net (PN) in a systems biology approach. Our focus is to model the interactions and regulatory processes of signalling to NF-κB, which are not described in a mathematical model so far. Recently gained insights like linear ubiquitylation events are considered in the model [Wal11]. We apply P/T-Petri nets to qualitatively model NF-κB signalling according to the literature, since no quantitative data is available. The MonaLisa tool is used for the construction, analysis, and simulation of the model [EANK13]. The mathematical analysis of the PN confirms the consistency of the model by fulfilling the CTI property. Analysis of invariants points out regulatory effects, and PN simulation reveals dynamics. The established PN reflects the current understanding of signalling to NF-κB, while delivering a valuable basis for further quantitative mathematical analysis.

## References

[EANK13] Jens Einloft, Jörg Ackermann, Joachim Noethen, and Ina Koch. MonaLisa - visualisation and analysis of functional modules in biochemical networks. *Bioinformatics*, 29:1469–1470, 2013.

[Wal11] Henning Walczak. TNF and ubiquitin at the crossroads of gene activation, cell death, inflammation, and cancer. *Immunological Reviews*, 244:9–28, 2011.

# Comparison of different graph-based pathway analysis methods on breast cancer expression data

Bayerlova M.[1], Kramer F.[1], Jung K.[1] , Klemm F.[2], Bleckmann A.[1,2], Beissbarth T.[1]

[1]*Department of Medical Statistics and* [2]*Department of Hematology/Oncology, University of Göttingen, 37099 Göttingen, Germany*
Michaela.Bayerlova@med.uni-goettingen.de

Pathway analysis methods are a frequently used bioinformatics approach to test enrichment or overrepresentation of differentially expressed genes in a given pathway. Probably the most popular among these methods is Gene Set Enrichment Analysis (GSEA). However, the main drawback of GSEA is that it handles pathways as gene lists omitting any knowledge of molecular interactions or graph structure. Recently, several bioinformatics algorithms integrating pathway topology information into pathways analysis were proposed: SPIA (Tarca *et al.* 2009), GGEA (Geistlinger *et al.* 2011) and clipper (Martini *et al.* 2012). We have compared and evaluated these graph-based pathway analysis methods using simulated expression data and graph information from different pathway databases (Pathway interaction database, KEGG, Reactome, Biocarta). To integrate graph data into our analysis we have developed an R package rBiopaxParser (Kramer *et al.* 2012) which imports BioPAX encoded pathway data into the statistical computing environment of R. Within R the parser is transforming data into a directed graph or an adjacency matrix and allows further editing of pathways to obtain suitable input for the before mentioned pathways analysis methods. Further, we have applied these algorithms to breast cancer expression data to test significance of WNT signalling pathways and sub-pathways parsed from several databases within the context of different molecular subtypes of breast cancer.

# NOVA: Evaluation of complexome profiling data

Heiko Giese [1], Jörg Ackermann [1], Heinrich Heide [2,3], Ilka Wittig [2,3], Ulrich Brandt [2,4], Ina Koch [1]

[1] *Molecular Bioinformatics Group, Institute of Computer Science, Faculty of Computer Science and Mathematics, Cluster of Excellence Frankfurt "Macromolecular Complexes", Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main, Germany*

[2] *Molecular Bioenergetics Group, Medical School, Cluster of Excellence Frankfurt "Macromolecular Complexes", Goethe-University, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany*

[3] *Functional Proteomics, SFB815 core unit, Medical School, Goethe-University, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany*

[4] *Nijmegen Centre for Mitochondrial Disorders, Radboud University, Nijmegen Medical Centre, Geert Grooteplein Zuid 10, NL-6500 Nijmegen, The Netherlands*

**Motivation:** The isolation of large native macromolecular complexes, identification of components and their dynamics, is a difficult task, requiring advanced proteomic strategies like complexome profiling [HBS+12]. Complexome profiling uses blue-native electrophoresis (BNE) to separate protein mixtures. Proteins that are subunits of the same complex are expected to have similar migration profiles which are measured by label-free quantitative mass spectrometry. Because manual comparison of all migration profiles is not feasible, cluster analysis is required to process the data sets. To the best of our knowledge no tool is available that offers statistical methods to evaluate complexome profiling data.

**Results:** We developed NOVA - a new tool for the analysis of complexome profiling data. A graphical user interface (GUI) provides various visualization modes, such as heat maps and 2D plots. Several hierarchical clustering algorithms (e.g., average linkage, Wards linkage), different distance measures (e.g., Euclidean distance, Pearson distance), and various normalization techniques are implemented. Migration profiles of several complexome profiling experiments can be easily compared (e.g. knock-down vs. wild typ). Many additional functions like zooming, searching for proteins, image export, and automatic file format recognition support intuitive handling. We demonstrate the functionality of the program by its application to recent experimental data obtained by complexome pro-

filing.

## References

[HBS+12] Heinrich Heide, Lea Bleier, Mirco Steger, Jörg Ackermann, Stefan Dröse, Bettina Schwamb, Martin Zörnig, Andreas S. Reichert, Ina Koch, Ilka Wittig, and Ulrich Brandt. Complexome Profiling Identifies TMEM126B as a Component of the Mitochondrial Complex I Assembly Complex. *Cell Metabolism*, 16(4):538–549, October 2012.

# Local Search for Bicriteria Multiple Sequence Alignment

Maryam Abbasi[1], Luís Paquete[1], Francisco Pereira[1,2] and Sebastian Schenker[3]
[1]*CISUC, University of Coimbra, Portugal*
[2]*Polytechnic Institute of Coimbra, Portugal*
[3]*Zuse Institute Berlin, Germany*

{maryam,paquete,xico}@dei.uc.pt, schenker@math.tu-berlin.de

Recently, there has been a growing interest on the multicriteria formulation of optimization problems that arise in computational biology, such as sequence alignment [T10,A13,S13]. In this work, we consider the multicriteria multiple sequence alignment, where the goal is to maximize the substitution score and minimize the number of indels or gaps. We introduce local search algorithms for several variants of this problem. The acceptance criterion of the local search is based on the dominance criterion [P07]. Several neighbourhood definitions and perturbations are presented and discussed. The local search algorithms are tested experimentally on a wide range of instances. The solution quality of this approach is compared against bounds obtained by solving a sequence of integer linear programming formulations by a known branch-and-cut approach for multiple sequence alignment [A06].

This work was supported by the Portuguese Foundation for Science and Technology and FEDER, Programa Operacional Factores de Competitividade - COMPETE, FEDER - FCOMP-01-0124-FEDER-010024, under the project "Multiobjective Sequence Alignment" (PTDC/EIA-CCO/098674/2008).

## References

[A06] E. Althaus, A. Caprara, H.-P. Lenhof, K. Reinert, A branch-and-cut algorithm for multiple sequence alignment, *Mathematical Programming, Ser. B*, 105:387-425, 2013

[A13] M. Abbasi, L. Paquete, A. Liefooghe, M. Pinheiro, P. Matias, Improvements on bicriteria pairwise sequence alignment: algorithms and applications, *Bioinformatics*, 29(8):996-1003, 2013

[P07] L. Paquete, T. Schiavinotto, T. Stützle, On local optima in multiobjective combinatorial optimization problems, *Annals of Operations Research*, 156(1):83-97, 2007

[S13] T. Schnattinger, U. Schöning, H. Kestler, Structural RNA alignment by multi-objective optimization, *Bioinformatics*, 29(13):1607-1613, 2013

[T10] A. Taneda, Multi-objective pairwise RNA sequence alignment, *Bioinformatics*, 26(19):2383-2390, 2010

# Boolean network reconstruction to explain individual drug response in breast cancer

*Silvia von der Heyde[1], Christian Bender[2], Frauke Henjes[3], Johanna Sonntag[4], Ulrike Korf[4], Tim Beißbarth[1]*

*1 Medical Statistics, University Medical Center Göttingen, Germany*

*2 TRON, University Medical Center Mainz, Germany*

*3 Science for Life Laboratory, KISP, Solna, Sweden*

*4 Molecular Genome Analysis, DKFZ, Heidelberg, Germany*

silvia.heyde@med.uni-goettingen.de

Breast cancer, a heterogeneous disease, requires individual therapies. Despite promising research on targeted therapeutics in personalized medicine, drug resistance remains challenging. The 'HER2-enriched' molecular subtype over-expresses the ErbB2-receptor, accounting for 10-20% of breast tumours. Amplifications or mutations of epidermal growth factor receptors (ErbB) can induce oncogenic protein signalling. The drugs erlotinib, trastuzumab and pertuzumab target ErbB1- and ErbB2-receptors, but heterodimerization, also with ErbB3, can bypass inhibition. We analyzed individual drug response and potential resistance mechanisms in three ErbB2-amplified breast cancer cell lines with different phenotypes, namely BT474, SKBR3 and HCC1954. The latter harbours a PI3K-mutation and is trastuzumab resistant. SKBR3 and HCC1954 also highly express ErbB1. Based on time-resolved phosphoproteomic reverse phase protein array (RPPA) data of ErbB-receptors and downstream targets under drug treatment, we reconstructed cell line specific signalling networks in a Boolean modelling approach, including prior literature knowledge [Ben10]. The inferred protein wiring revealed individual cell line characteristics with different pathway preferences and potential resistance mechanisms. Subsequently, we simulated drug responses of these networks to detect tumour suppressing inputs, reflected by deactivated perturbation related stable system states of MAPK and PI3K pathway targets [Müs10]. Novel insights into phenotype specific dysregulation in the ErbB-network can be integrated into predictive personalized drug response models.

## References

[Ben10] Bender et al. Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data. *Bioinformatics*, 26(18):i596-602, 2010

[Müs10] Müssel et al. BoolNet--an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*, 26(10):1378-80, 2010

# Modeling and Simulation of Biological Networks using extended hybrid functional Petri nets

Christoph Brinkrolf, Sebastian Jan Janowski, Lennart Ochel, Martin Lewinski, Benjamin Kormeier, Bernhard Bachmann and Ralf Hofestädt
*AG Bioinformatik / Medizinische Informatik, Technische Fakultät, Universität Bielefeld*
cbrinkro@cebitec.uni-bielefeld.de

During the process of creating models of biological networks one often comes to the point of lacking information. At this point one is no longer able to model quantitatively. Therefore, a formalism is necessary that allows to create qualitative, quantitative and combined models and enrich these models successively with additional information. Furthermore, the formalism should be qualified for describing the biological system in detail. This can be done by Petri nets, first defined by Carl Adam Petri [Pet62], and their extensions. The basic definition containing only discrete elements has been extended in the following years by concepts of e.g. continuous elements, capacities, functions, inhibitory edges, hierarchical structures and colored tokens. For a transparent and correct simulation extended hybrid Petri nets (xHPN) have to be well-defined, which is done in [PB12] and [Pro13]. In our tool VANESA [Jan13], a qualitative model can be retrieved from our data warehouse DAWIS-M.D. [HKT$^+$10] where several biological databases are integrated. Quantitative data can be added using again DAWIS-M.D. or data from laboratory experiments. For the process of simulation, the biological network is converted into a fully editable Petri net. The simulation itself is realized in a Petri net library called PNlib [PB12, Pro13], which, is an implementation of xHPN in the object oriented modeling language Modelica. Graph theory and Petri net theory in particular provide a variety of established analysis techniques that are well-suited and applicable to biological network modeling. These analyses and the analysis of simulation results then will lead to new observations, hypotheses, more precise models and a deeper understanding of the investigated biological system.

## References

[HKT+10] Klaus Hippe, Benjamin Kormeier, Thoralf Töpel, Sebastian Jan Janowski, and Ralf Hofestädt. DAWIS-M.D. - A Data Warehouse System for Metabolic Data. *GI Jahrestagung (2) 2010: 720-725*, 2010.

[Jan13] Sebastian Jan Janowski. VANESA - A bioinformatics software application for the modeling, visualization, analysis, and simulation of biological networks in systems biology applications. *Dissertation, Universität Bielefeld, Technische Fakultät*, 2013.

[PB12] Sabrina Proß and Bernhard Bachmann. PNlib An Advanced Petri Net Library for Hybrid Process Modeling. *Modelica Conference*, 2012.

[Pet62] Carl Adam Petri. Kommunikation mit Automaten. *Dissertation, Rheinisch-Westfälisches, Institut für Instrumentelle Mathematik*, 1962.

[Pro13] Sabrina Proß. Hybrid Modeling and Optimization of Biological Processes. *Dissertation, Universität Bielefeld, Technische Fakultät*, 2013.

# Detection of synergistic effects evoking new functions in a cell using a bipartite network algorithm

*Sebastian Zeidler, Björn Goemann, Raphaël Zollinger* and Edgar Wingender*

*Institute of Bioinformatics, University Medical Center Göttingen*
**Institut Curie, Laboratoire d'Immunologie Clinique*

Simultaneously acting signals affect the behavior of organisms at systems level. Different signals must be integrated and evaluated simultaneously. Synergistic interactions are one possibility how a cell integrates different signals. By this, qualitatively new functions may be induced. Experimental studies observing this type of synergy are rare, due to the high complexity of combinatorial testing. Gene expression approaches detect affected genes, but the underlying mechanisms remain unclear. We present an algorithmic approach for the identification of synergies in signal transduction networks using a reconstructed bipartite network containing mechanistic information from the TRANSPATH® and TRANSFAC® databases [1,2]. The separation of vertices into two subsets (biological entities, signal transduction processes) enables the detection of synergies as topological patterns. Synergistic signal integration happens during exclusively activated processes. The algorithm has been applied to an immune framework to identify synergistic effects induced by IL-3 and Flu in plasmacytoid dendritic cells. Mammals possess an elaborated immune where synergistic effects are frequently observed [3]. We detect synergistic signal integration in the singalling pathways and predict downstream affected genes. We show that the algorithmic approach detects immune relevant synergy more precisely than the statistical analysis of experimental data.

**References**:

[1] C. Choi et al. Consistend Re-Modeling of Signaling Pathways and Its Implementation in the TRANSPATH Database, Genome Inform., 15:244-254, 2004.

[2] E. Wingender, The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation, Brief. Bioinform. 9:326-332,2008.

[3] M. Gouwy et al. Synergy between proinflammatory ligands of G protein-coupled receptors in neutrophil activation and migration, J. Leukoc. Biol. 76:185-194, 2004.

# The $k$-Mismatch Average Common Substring approach

Chris Leimeister and Burkhard Morgenstern
*University of Göttingen, Department of Bioinformatics, Goldschmidtstr. 1, 37077 Göttingen, Germany*
Chris.Leimeister@stud.uni-goettingen.de

Alignment-free methods for phylogeny reconstruction have recently become popular because they are much faster than traditional alignment-based methods. Some alignment-free methods are based on relative entropy such as the average common substring approach[UBTC06]. To define the similarity between two sequences, this approach computes for each position $i$ in the first sequence the length of longest substring starting at $i$ and matching some substring of the second sequence. It defines the average of these values as a measure of similarity between the sequences and turns this into a symmetric distance measure.

Herein, we suggest to generalize this approach by considering the longest substring starting at position $i$ in the first sequence and matching a substring in the second sequence with $k$ mismatches. To approximate the length of the longest such substring, we first consider the longest *exact* match starting at $i$ to some substring of the second sequence, and we then extend this match without gaps until we reach the $k+1$-th mismatch.

To get a first impression of the performance of this approach, we used simulated families of DNA and protein sequences and constructed phylogenetic trees based on our new distance measure. These trees were compared to the known reference trees for the respective sequence sets. We found out that our approach leads to far better results than the average common substring method without mismatches. Increasing the value of $k$ generally improves the resulting trees. The improvement is most significant for small values of $k$; for larger values, the quality of the trees tends to converge.

## References

[UBTC06] Igor Ulitsky, David Burstein, Tamir Tuller, and Benny Chor. The Average Common Substring Approach to Phylogenomic Reconstruction. *Journal of Computational Biology*, 13:336–350, 2006.

# High Betweenness – Low Connectivity (HBLC) Signatures in the Human Proteome

Thomas Wiebringhaus and Heinrich Brinck
*Institute of Bioinformatics and Chemoinformatics,*
*Westphalian University of Applied Sciences*
thomas.wiebringhaus@w-hs.de

One major question in current network biology is the decomposition of cellular behaviour into smaller biological functions to understand the complex interwoven web of protein interactions. It is a well-known condition that modules have to be connected to exchange information and thus have to be linked via intermodular links or mediating complexes.

By applying established graph theory methods, a new interesting topological feature was recently discovered for yeast protein interaction networks by calculating the *betweenness centrality (BC)* (Joy et al., 2005). The *BC* measures how central a specific network component is for the overall communication of the network. Although this attempt is only technically meaningful here, as the calculated shortest paths are not necessarily biological pathways, the method might be practicable to get new insights for topological features.

It is found here that the distribution pattern of the *BC* as a function of the *degree* is very similar to yeast, suggesting a biological basis. The finding is not explained by typical scale-free properties of networks, where *low-connectivity* proteins also have *low-betweenness centrality* (Nakao, 1990; Goh et al., 2003). It was supposed by Joy et al. that these *HBLC* proteins might represent functionally important proteins that act as intermodular connections. In contrast to previous results for yeast, the found *HBLC* proteins are biologically interpreted here for the human proteome.

## References

Gavin AC et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, 415(6868) 141-7, 2002

Goh KI et al. (2003) Betweenness centrality correlation in social networks, *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(1 Pt 2): 017101

Joy MP et al. (2005) High Betweenness proteins in the yeast protein interaction network, *J Biomed Biotechnol*, 2005(2), 96-103

Nakao K. (1990) Distribution of measures of centrality: Enumerated distributions of Freeman's graph centrality measures, *Connections*, 13(3), 10-22

# An efficient approach to generate chemical substructures for MS/MS peak assignments in MetFrag

Christoph Ruttkies and Steffen Neumann
*Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany*
{cruttkie|sneumann}@ipb-halle.de

MetFrag [1] is a tool for the annotation of tandem mass (MS/MS) spectra via *in silico* fragmentation of small molecules, in particular for identification workflows in Metabolomics. The fragmentation process of selected molecules to assign fragments to mass peaks in given MS/MS spectra is a crucial step. The quickly growing metabolite databases require reduced computing time as well as less memory usage to process candidate molecules because MetFrag previously used a brute force method to generate fragment substructures.

Due to the representation of molecules as graphs – where vertices represent the atoms and edges represent the bonds – we can benefit from existing algorithms to generate subgraphs representing the fragments of the given molecule.

We present a combined approach based on *Community* calculation (provided by the R package igraph [2]) and *Subgraph Enumeration (ESU)* [3]. *Communities* representing subgraphs share common properties with fragments of a molecule generated by low-energy tandem mass spectrometry. Within each highly connected *Community*, all fragments are enumerated by ESU. Then, we combine in a brute force manner any number of *Communities* with zero or one ESU generated fragment from another *Community*.

This subdivision into *Communities* effectively reduces the number of nodes and edges and reduces runtime and memory requirements depending on the *Community* sizes. The evaluation of our benchmark data consisting of 1099 MS/MS spectra showed a better performance of the new approach. Moreover, the rankings of the correct candidates could be improved.

## References

[1] Wolf, S. *et al. BMC Bioinformatics* **2010**, *11*, 148.

[2] Clauset, A. *et al. Physical Review E* **2004**, *70*, 066111.

[3] Wernicke, S. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* **2006**, *3*, 347–359.

# Gamification of gene prediction

Klas Hatje, Dominic Simm and Martin Kollmar

*Max Planck Institute for Biophysical Chemistry, Göttingen, Germany*
hakl@nmr.mpibpc.mpg.de

The availability of next generation sequencing methods has led to an exponential growth in the number of sequenced eukaryotic genomes [Ham11]. In contrast to this progress, advances in the process of genome annotation (in the sense of gene finding) lags behind. Even though bioinformatic methods are used routinely for genome annotation no experimental or bioinformatic solution exists that is exact and comprehensive. Here, we present a new approach to generate correct, reliable and complete annotations of all genes in eukaryotes with the aid of human problem-solving strategies. In the form of an online computer game we hope to obtain voluntary contributions of people outside the traditional scientific community. Building on the annotation of the human genome, related mammalian genomes will be annotated based on reconstructed genes provided by Scipio [Hat11]. The task of the players is to complete the annotations with the help of multiple protein and DNA sequence alignments, phylogenetic information and gene structure comparisons.

## References

[Ham11] Hammesfahr B, Odronitz F, Hellkamp M, Kollmar M. diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data. *BMC Research Notes*, 4(338), 2011.

[Hat11] Hatje K, Keller O, Hammesfahr B, Pillmann H, Waack S, Kollmar M: Cross-species protein sequence and gene structure prediction with fine-tuned Webscipio 2.0 and Scipio. *BMC Research Notes*, 4(265), 2011.

# Identification of gene co-expression networks associated with different cellular and immunological states

Marc Bonin[1], Jekaterina Kokatjuhha[1], Stephan Flemming[2], Biljana Smiljanovic[1], Andreas Grützkau[3], Till Sörensen[1], Thomas Häupl[1]

[1]*Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin*
[2] *Institute of Pharmaceutical Sciences, University of Freiburg*
[3]*German Arthritis Research Center, Berlin*
marc.bonin@charite.de

**Introduction**
Knowledge about gene networks is of great importance for analysis of transcriptome data. However, current tools mainly rely on information about direct molecular interactions between proteins, which is not directly connected to expression levels. These differences between transcriptome based perception of biological information and tools for network analysis are the main reason for difficulties in functional interpretation. Therefore, we started to use transcriptome data of biologically well-defined states to define functional markers and signatures as tools for future analysis.

**Methods**
GeneChip HG-U133 Plus 2.0 transcriptomes from highly purified blood cell types (granulocytes, monocytes, CD4+ and CD8+ T-cell, B-cells, NK-cells) as well as from monocyte stimulation with LPS, TNF and type 1 IFN were selected from the BioRetis database (www.bioretis.de). Correlations of expression between all probesets were calculated to filter for co-regulation. Correlation matrices were calculated, clustered and displayed in heat maps. The web-platform www.humanresearchdb.charite.de was constructed based on Ruby on Rails to provide a framework for analysis and storage of data.

**Results**
Initially, correlation matrices were determined for each individual stimulation condition and its control. Stepwise combination of the three different conditions for calculation of correlation coefficients revealed a reduction of the correlation network and a reduction of overlap between the networks. This indicates increasing functional specificity of the identified candidates. All of the typical previously published IFN related genes were identified and thus confirmed our strategy. In a similar way, cell type specific co-expression networks were determined. Additional filtering for high signal intensity provides candidates for sensitive detection of the function related patterns even in highly diluted conditions. These marker panels are currently tested for detection and quantification of functional signatures in biopsies of inflamed tissue.

**Conclusion**
Correlating transcription between genes in well-defined biological states identifies function-related markers and signatures. Depending on the type of function, appropriate conditions have to be selected.

# Combining features for protein interface prediction

Torsten Wierschin[1], Keyu Wang[2], Stephan Waack[3], Mario Stanke[4]
*Institut für Mathematik und Informatik, Universität Greifswald*[1,4]
{torsten.wierschin, mario.stanke}@uni-greifswald.de
*Institut für Informatik, Universität Göttingen*[2,3]
{kwang,waack}@informatik.uni-goettingen.de

Interface prediction can be formulated as a problem in which a binary classsification (interface or not) of all residues of one protein with a given 3D structure is sought. Previous work includes the independent classification of each residue with machine learning approaches. Li et al. considered some of the interdependencies between labels of different residues by interpreting the task as a sequential labeling problem and using a linear-chain conditional random field (CRF) [LLWL07]. In contrast, our CRF makes the weaker 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 within threshold distance $(3\mathring{A} - 12\mathring{A})$. We consider linear combinations of various feature functions previously defined in the literature. We introduce the new feature change in free energy: an *in silico* mutation of a residue changes the free energy of the whole protein considerably if the residue is indeed an interface. We achieve more accurate results than the SVM based *PresCont* server [ZST+11].

## References

[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.

[ZST+11] H. Zellner, M. Staudigel, T. Trenner, M. Bittkowski, V. Wolowski, C. Icking, and Rainer Merkl. Prescont: Predicting protein-protein interfaces utilizing four residue properties. *Proteins*, September 2011.

# Circular permutations: detecting evolutionary related protein pairs based on structure analysis

Martin Mosisch[1], Thomas Margraf[2] and Andrew Torda[1]
[1]*Center for Bioinformatics Hamburg,* [2]*EMBL Hamburg*
mosisch@zbh.uni-hamburg.de

Circular permutated proteins reflect a genetic event in which part of the C-terminus has moved to the N-terminus or vice versa as if on a circle. Thus a protein ABCD may be related to BCDA or DABC. CPs are prevalently thought to arise via three mechanisms: (1) gene duplication/ deletion, (2) ”independent fusion” reflecting protein domain fusion and fission, or (3) the cut and paste mechanism of the restriction-modification (RM) system.[WB06] Such evolutionary events are considered to be rare, but their detection can help determine structural domains and designing recombinant proteins. Nonlinear rearrangements cause traditional dynamic programming methods based on sequence similarity to find only partial similarities. The consequence is that very distant evolutionary events will be missed. We use an extended alignment matrix to detect permuted relations, but perform structural instead of sequence comparisons. Structure comparison is based on probabilities of fragment-membership from an earlier structural classification.[MGT09] Our method is not free of thresholds and adjustable parameters, but as set now, we find 95% of previously documented examples. In order to process a representative large scale test set while bypassing a full scan of all chains in the protein database, firstly, we discretized structure space by encoding probability vectors for protein chains to structure probability sequences and formed a suffix tree, secondly, through querying this suffix tree we clustered protein chains into groups by their k-nearest neighbors (~18 Mio. pairs). Our method applied to this large set discovered some remarkable evolutionary relations within an appropriate period, amongst others that CPs are not that infrequent as one would normally expect.

## References

[MST09] Thomas Margraf, Gundolf Schenk and Andrew E. Torda. The SALAMI protein structure search server. *Nucleic Acids Research*, 37**:**W480–W484, 2009.

[WB06] January Weiner and Eric Bornberg-Bauer. Evolution of Circular Permutations in Multidomain Proteins. *Molecular Biology and Evolution*, 23(4): 734–743, 2006

# A scalable method for the correction of homopolymer errors

Giorgio Gonnella and Stefan Kurtz
*Center for Bioinformatics (ZBH), University of Hamburg*
gonnella@zbh.uni-hamburg.de

Error correction is an important preprocessing step for next-generation sequencing data. Most error correction algorithms, such as existing $k$-mer based correction methods, have been designed for substitution errors. However, in some widely available sequencing platforms, Roche 454 and IonTorrent, the dominant errors are insertions and deletions in homopolymers [LMD+12].

Despite this, no methods have been presented, which are specifically targeted to homopolymer errors. Instead, general-purpose alignment-based methods are usually applied to correct homopolymer errors too. However, the homopolymer-error prone platforms are steadily increasing their throughput, so that alignment-based methods will soon require too many computational resources for processing an entire dataset.

Here we propose a new $k$-mer based algorithm for the correction of homopolymer errors. Our method is effective and scalable to high-coverage datasets: with a proof-of-concept implementation, we were able to correct a simulated Roche 454 *E.coli* dataset with 640× coverage in 14.1 CPU hours using 3.3 Gb RAM and achieving 98.55% sensitivity and 77.29% specificity. On the same dataset, the state-of-the-art alignment-based correction tool Coral [SS11] requires 262.5 CPU hours and 66.6 Gb RAM, achieving a sensitivity of only 84.42% and specificity of 35.57% while correcting homopolymers.

## References

[LMD+12] Nicholas J Loman, Raju V Misra, Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, John Wain, and Mark J Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30(5):434–9, May 2012.

[SS11] Leena Salmela and Jan Schröder. Correcting errors in short reads by multiple alignments. *Bioinformatics*, 27(11):1455–61, June 2011.

# Multiple genome comparison based on overlap regions of pairwise local alignments

Katharina Jahn, Henner Sudek, and Jens Stoye

Technische Fakultät, Universität Bielefeld, Germany,
kjahn@cebitec.uni-bielefeld.de, stoye@techfak.uni-bielefeld.de

Comparative approaches are an important source of information when it comes to the analysis of newly sequenced genomes. On the level of genes, the use of reciprocal BLAST hits is the most widely accepted approach suitable for tasks like gene annotation and the inference of homologies. However, it is a notoriously slow process, especially when it comes to all-against-all comparisons in a large amount of genomes.

Recently, Mancheron *et al.* [1] introduced a new approach in the context of multiple genome comparison that allows to detect regions of strong overlaps in a set of pairwise local alignments between several reference genomes and one target genome. Such overlap regions are an important source of information for the transfer of genome annotations.

We introduce a series of algorithms that improve over the approach of Mancheron et al., both in terms of computational complexity and in practical runtime. We also extend the problem definition such that overlaps to different reference genomes can be rated differently and regions overlapping only a subset of the reference genomes are detected [2]. We also study a variant of this problem where pairwise overlaps are weighted individually which allows to consider pairwise alignment scores in the assessment of an overlap region.

## References

1. Mancheron A, Uricaru R, Rivals E: **An alternative approach to multiple genome comparison**. *Nucleic Acids Res.* 2011, **39**:e101.
2. Jahn K, Sudek H, Stoye J: **Multiple genome comparison based on overlap regions of pairwise local alignments**. *BMC Bioinformatics* 2012, **13**:S7(Suppl. 19)

# Enrichment Analysis for Hierarchical Clusters

J.T. Kim[1*], K. Staines[1], J. Young[1], Z. Minta[2], K. Smietanka[2], D. Balkissoon[1], R. Ruiz-Hernandez[1] and C. Butter[1]
[1] *The Pirbright Institute, Woking GU24 0NF, United Kingdom*
[2] *National Veterinary Research Institute (PIWET), Poland*
[*] jan.kim@pirbright.ac.uk

Transcriptomic experiments are routinely used to identify candidate genes involved in mediating responses to biological signals or environmental challenges. Ranking genes using linear models, and ontology (GO) term enrichment analysis, and hierarchical clustering are useful classification tools to support further biological interpretation of such candidate gene sets.

We present a workflow which combines these tools by identifying a set of top ranking genes using limma [Smy04], structuring this set using hierarchical clustering, and applying enrichment analysis using topGO [ARL06] to each cluster. This enables identification of clusters in which enrichment is strongest, and can thereby aid in more systematically targeting genes for further investigation. The entire workflow is implemented in R [R D04].

We demonstrate application of the workflow to gene expression data from chicken embryo fibroblast cultures. Cells were infected by influenza, a challenge which results in massive changes of gene expression levels, with or without pre-treatment with interferon $\alpha$ (IFN$\alpha$), a signal that induces antiviral responses.

## References

[ARL06] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure. *Bioinformatics*, 22:1600–1607, 2006.

[R D04] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.

[Smy04] Gordon K. Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article 3, 2004.

# A general approach for discriminative de-novo motif discovery from high-throughput data

Jan Grau [1], Stefan Posch [1], Ivo Grosse [1], and Jens Keilwagen [2]
[1] *Institute of Computer Science, Martin Luther University Halle–Wittenberg, Halle*
[2] *Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg*
grau@informatik.uni-halle.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 from data obtained by wet-lab experiments is still a challenging problem in bioinformatics, and has not been fully solved yet. Today, major sources of *in-vivo* and *in-vitro* data are chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-seq/ChIP-exo) and protein binding microarrays (PBMs), respectively.

We present Dimont, a de-novo motif discovery approach specially tailored to these high-throughput data. Dimont successfully discovers all motifs of the ChIP-seq data sets of Ma *et al.* [M+12]. On the data sets of Weirauch *et al.* [W+13], it predicts PBM intensities from probe sequence with higher accuracy than any of the approaches specifically designed for that purpose. Dimont also reports the expected motifs for several ChIP-exo data sets. Investigating differences between *in-vitro* and *in-vivo* binding, we find that for most transcription factors, the motifs discovered by Dimont are in good accordance between techniques, but we also find notable exceptions. We provide a Dimont web-server at `http://galaxy.informatik.uni-halle.de` and a command line application at `http://www.jstacs.de/index.php/Dimont`.

## References

[M+12] Xiaotu Ma et al. A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res*, 40(7):e50, 2012.

[W+13] Matthew T. Weirauch et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotech*, 31:126–134, 2013.

# Novel Visualization Approach Integrating Network and Structure Analysis of Proteins

Nadezhda T. Doncheva[1], John H. Morris[2], Eric F. Pettersen[2], Conrad C. Huang[2], Karsten Klein[3], Francisco S. Domingues[4], Thomas E. Ferrin[2], Mario Albrecht[1,5]

[1]*Max Planck Institute for Informatics, Saarbrücken, Germany*
[2]*University of California, San Francisco, USA*
[3]*University of Sydney, Sydney, Australia*
[4]*EURAC research, Bolzano, Italy*
[5]*University Medicine Greifswald, Greifswald, Germany*

nadezhda.doncheva@mpi-inf.mpg.de

To understand complex molecular mechanisms, combining systems biology and structural biology will be very beneficial for both fields [FrGK13]. Our recent work already linked the visualization of biological networks in Cytoscape to the visualization and analysis of protein structures in UCSF Chimera [MHBF07]. This included a novel approach to the interactive visual analysis of residue interaction networks derived from 3D protein structures [DKDA11]. Here, we present substantial extensions of our tools structureViz and RINalyzer, which are now integrated even better with the new releases of Cytoscape and UCSF Chimera. By enhancing molecular networks with sequence and structure information and by providing a complementary network representation of residue interactions, our tools facilitate a novel interactive, multi-layered analysis of protein interactions and their molecular function in protein binding, allosteric effects, drug resistance and other mechanisms. Additionally, we introduce a new approach for visualizing and analyzing ensembles of protein structures as generated in molecular dynamics by representing them as networks of probabilistic residue interactions.

**References**

[FrGK13] J.S. Fraser *et al*. From systems to structure: bridging networks and mechanism. *Mol Cell*, 49(2):222-231, 2013.

[MHBF07] J.H. Morris *et al*. structureViz: linking Cytoscape and UCSF Chimera. *Bioinformatics*, 23(17): 2345-7, 2007.

[DKDA11] N.T. Doncheva *et al.* Analyzing and visualizing residue networks of protein structures. *Trends Biochem Sci.*, 36(4):179-182, 2011.

# Protein Subcellular Location Prediction Using Principal Component Analysis

Daichi Nogami,Yuichi Nakano and Y-h. Taguchi

*Chuo University*

*daichi.n45@gmail.com*

It is known that subcellular location of proteins is important for elucidating their functions involved in various cellular processes. Till now,many efforts for predicting the subcellular localization more accurately have been tried. Especially, Inference based upon only amino acid sequence was mostly useful. These researches said that inference of subcellular location only using sequence information is important. In our research,we used simpler method for inference of subcellular location. We tried to predict the subcellular location of proteins using principal component analysis of the physicochemical features calculated from protein amino acid sequence and confirmed that it is an effective means. And we proposed the application of principal component analysis-based linear discriminant analysis for subcellular location information. Our research's performance measure are Matthews Correlation Coefficient and Area Under the Curve under Receiver Operating Characteristic curve. We tried to compare performances between our method and GO term based method. Then we found that our sequence method can compete with GO term based method if the numbers of proteins in positive and negative sets do not differ from each other so much. It achieved the performance of Area Under the Curve under Receiver Operating Characteristic curve mostly more than 0.95 if less than 80% sequence non- redundancy was applied.

# Analyzing taxon and pathway coverage profiles with applications to metatranscriptomics

Daniela Beisser and Sven Rahmann
*Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, Germany*
daniela.beisser@uni-due.de

High-throughput sequencing has recently attracted much interest in the field of biodiversity research and has led to an increasing number of studies focusing on complex metagenomic and metatranscriptomic questions. Environmental factors influence on the one hand species richness and composition, on the other hand the functional repertoire of the whole ecosystem. To investigate both aspects, novel bioinformatic methods are needed to accurately and efficiently assign taxonomic as well as functional annotation to each sequenced short read.
This study analyses metatranscriptomes under changing environmental conditions and investigates adaptation of populations and functions. The annotation process consists of two main parts, the assignment of reads (I) to taxa and (II) to genes and metabolic pathways. Due to the lack of an adequate, unbiased database for taxonomic assignment, a reference database was built consisting of sequences from a selected set of species from all domains of life. The reads are mapped subsequently against the database in a two-step approach to obtain higher taxonomic orders (e.g. phylum). Information on gene functions and associations to the KEGG metabolic pathways are retrieved by mapping the reads to the UniProtKB database, using RAPSsearch2 [ZTY12]. Comparative analyses between environmental conditions, e.g. salt concentration or heavy metal contamination, are then performed on basis of the sample-specific taxon coverage and pathway profiles. For the statistical analysis appropriate normalizations are applied to account for differences in sequence lengths, uneven number of reads per sample and completeness of sequence information per taxon.

## References

[ZTY12] Yongan Zhao, Haixu Tang, and Yuzhen Ye. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126, Jan 2012.

# Bringing together only what belongs together: Characterizing and distinguishing protein structure families using distances based on contact map overlap

Inken Wohlers[1], Gunnar W. Klau[2] and Rumen Andonov[3]
[1]*Genome Informatics, University of Duisburg-Essen, Germany*
[2]*Life Sciences, Centrum Wiskunde & Informatica, the Netherlands*
[3]*INRIA Rennes - Bretagne Atlantique, Rennes, France*
inken.wohlers@uni-due.de

Contact map overlap is a commonly accepted simple, yet powerful key figure for protein structure similarity. Given an alignment, it is defined as the number of contacts whose endpoint residues are aligned. The NP-hard problem of finding an alignment with maximum contact map overlap can be considered as a graph matching problem, for which several graph distances have been proposed. These distances can serve as measures for comparing protein structures. Here we present such measures and use them to compute all pairwise distances between protein family members from a SCOP and CATH consensus benchmark [CBZ09] by applying our exact contact map overlap solver [AMDY11]. Distances are used to determine a central, representative structure for each family as well as a radius around it which contains all family members. Using this radius as well as inter-family distances we observe the characteristics of each family, for example its homogeneity, variability and possible uniform subgroups. Further we compute the distances between family representatives to investigate which and how many families overlap and to which extend. By doing so, we evaluate the performance of different distance measures with respect to pooling similar and distinguishing dissimilar structures.

## References

[AMDY11] R. Andonov, N. Malod-Dognin, and N. Yanev. Maximum contact map overlap revisited. *J Comput Biol*, 18(1):27–41, 2011.

[CBZ09] G. Csaba, F. Birzele, and R. Zimmer. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Struct Biol*, 9:23–23, 2009.