Central German Meeting on Bioinformatics 2015

Abstracts

Martin Luther University Halle–Wittenberg
August 26/27, 2015

Organizing institutions

MARTIN-LUTHER-UNIVERSITÄT
HALLE-WITTENBERG

U.Z.I Universitätszentrum
Informatik - Halle

IPB Leibniz Institute of
Plant Biochemistry

iDiv German Centre
for Integrative
Biodiversity Research

IPK
GATERSLEBEN

Supporting institutions & Sponsors

ScienceCampus
PLANT-BASED BIOECONOMY
HALLE

clustervision

ontochem
IT SOLUTIONS

DELL
The power to do more

MEGWARE®
SUPERCOMPUTING · TECHNOLOGY

GOVIRTUAL
Germany GmbH

sgi®

Keynotes

Francios Buscot

Helmholtz Centre for Environmental Research - UFZ, Halle
German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig

New perspectives and challenges in plant-soil ecology research at the age of next generation sequencing

Nicole van Dam

Institute for Ecology, Friedrich Schiller University Jena
German Center of Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig

Managing multiple menaces: Transcriptomics and metabolomics analysis of multiple stress responses in plants

Rainer König

Center for Sepsis Control and Care, Jena University Hospital

Inferring sample specific regulators for the genes of interest using gene expression data and chromatin binding information

Invited Talks

Stefan Schuster

Dept. of Bioinformatics
Friedrich Schiller University Jena
Faculty of Biology and Pharmacy

Abstract:

The Fibonacci code in lipidomics

Stefan Schuster[1], Severin Sasso[2]

[1] *Dept. of Bioinformatics, Friedrich Schiller University, Ernst-Abbe-Platz 2, 07743 Jena, Germany*
[2]*Institute of General Botany and Plant Physiology, Friedrich Schiller University, Dornburger Str. 159, 07743 Jena, Germany*

Fatty acids are of crucial importance for all living beings. In lipid biochemistry, lipidomics, and many other fields, an important question is how the potential number of fatty acids increases with their chain length. Here, we show that it grows according to the famous series of Fibonacci numbers when cis/trans isomerism is neglected. That series is defined by a recursive equation: each number is the sum of its two predecessors. Since the ratio of two consecutive Fibonacci numbers tends to the Golden section, 1.618, organisms can increase the variability of the fatty acids approximately by that factor per carbon invested into a fatty acid. The fraction of fatty acids with a terminal single bond is approximately given by the inverse of this ratio, 0.618. Under consideration of cis/trans isomerism, the numbers grow exponentially with the basis of two. We present some extensions of the calculations to modified fatty acids containing hydroxy and/or oxo groups. Depending on the group of molecules and isomers considered, different series such as the Pell numbers and generalized Tribonacci numbers are relevant. Our results should be of interest for mass spectrometry, combinatorial chemistry, synthetic biology, patent applications, the use of fatty acids as biomarkers and the theory of evolution.

Manja Marz

Bioinformatics/High Throughput Analysis
Friedrich Schiller University Jena
Faculty of Mathematics und Computer Science

Abstract:

Challenges in virus genomics

Computer-assisted studies of structure, function, and evolution of viruses remains a neglected area of research. The attention of bioinformaticians to this interesting and challenging field is far from commensurate with its medical and biotechnological importance. The purpose of this talk is to increase awareness among bioinformatics researchers about the pressing needs and unsolved problems of computational virology. I focus primarily on RNA viruses that pose problems to many standard bioinformatics analyses due to their compact genome organization, fast mutation rate, and low evolutionary conservation.

Steve Hoffmann

Transcriptome Bioinformatics
LIFE - Leipzig Research Center for Civilization Diseases
University Leipzig

Abstract:

Cancer Epigenomics - attempts to disentangle methylation, histone modification and transcription

Here, we present the integrative analysis of a whole genome bisulfite, genome and transcriptome sequencing from Burkitt lymphoma, follicular lymphoma and controls. Our investigation reveals regions of differential methylation (DMRs) that enrich transcription factor binding sites and are strongly correlated with the expression of associated genes. In the case of the transcription factor SMARCA4, our integrative analysis links this phenomenon to recurrent mutations in its helicase domain in Burkitt lymphoma. Furthermore, we report that genes under the control of poised (bivalent) chromatin are frequently hypermethylated and up-regulated at the same time. In an additional study we corroborate that hypermethylation of poised promoters and enhancers is a process common to a wide range of cancer types. Based on the relative methylation of originally poised regions, we propose a simple model to tell cancer samples from controls with high specificity and sensitivity. Finally, our pan-cancer analysis confirms that hypermethylation of formerly poised chromatin is frequently associated with up-regulation of associated genes.

Peter Stadler

Bioinformatics
University Leipzig
Institute of Computer Science

Abstract:

Orthology, Paralogy, and Cographs: How to use paralogs in phylogenomics

Recent advances in mathematical phylogenetics set the stage to actively making use of paralogous genes in phylogenomics application rather than spending efforts to exclude them as a nuisances. Orthology and (conversely) paralogy of genes can be estimated with reasonable accuracy from sequence similarity data. The orthology and paralogy relations mathematically are cographs. The empirical estimates thus can be corrected computationally to conform to the required mathematical structures. The resulting cographs in turn are equivalent to homeomorphic images of the gene trees together with an event labeling that assigns to each interior node whether it corresponds to a speciation or a gene duplication. Finally, one can show that the underlying species tree must diplay a certain subset of the triples defined by the event-labeled gene trees. Thus the orthology relation, or more precisely, the paralogs, convey independent information that can be used to successfully reconstruct the species tree itself.

Contributed Talks

e!DAL - A new approach to publish your data

Daniel Arend, Jinbo Chen, Christian Colmsee,
Uwe Scholz, Matthias Lange
*Leibniz Institute of Plant Genetics and
Crop Plant Research (IPK) Gatersleben*
daniel.arend@ipk-gatersleben.de

Research in the area of systems biology, phenotyping, and high-throughput technologies such as next-generation sequencing are generating a massive amount of data. This data explosion enables amazing discoveries, but also arise new demands for the data management and public access. Although several international consortia and public resource centre offer services for data maintenance and management, there are shortcomings in respect of sustainability, reproducibility, integration and usability. The consequence is a high risk to loose data that was not processed due to limited resources, subjectively classified as experimental garbage or considered not to be worth to put effort into its analysis or publication respectively. Estimations suggest a data loss rate of about 80% over 20 years [GVN13].

In this contribution we discuss experiences using the e!DAL framework [ALC$^+$14] to turn even small and mid-size institutions into public registered data centre. We developed, a handy data publication tool that provide an ad-hoc submission of research data and its registration at DataCite consortium as citable and documented scientific asset. Current data hotspot at IPK is the management of genomic and phenotypic data. We report experience from 6 month productive data publication service. So far we registered DOI's for over 50 datasets with a total volume of around 70 GB in about 21,000 files. In this period more than 80 TB was downloaded. This illustrates the public demand to get access to research data. The software, documentation and links to all productive, e!DAL powered data repositories are available at `http://edal.ipk-gatersleben.de`.

References

[ALC$^+$14] Daniel Arend, Matthias Lange, Jinbo Chen, Christian Colmsee, Steffen Flemming, Denny Hecht, and Uwe Scholz. e!DAL - a framework to store, share and publish research data. *BMC Bioinformatics*, 15(1), 2014.

[GVN13] Elizabeth Gibney and Richard Van Noorden. Scientists losing data at a rapid rate. *Nature*, 504, 2013.

Remolding and Evolution of tRNA genes in Eukaryotes

Sarah Berkemer
Bioinformatics Leipzig
bsarah@bioinf.uni-leipzig.de

It is known that tRNA gene and pseudogene loci undergo rapid evolution [BSSOAK$^+$10]. Based on orthology and synteny information, events such as gains, losses and pseudogenizations can be detected within a set of related species.
These events are partially caused by point mutations within the tRNA gene or duplications of the tRNA gene sequence within the genomic surroundings. Point mutations frequently lead to significant changes in the tRNA secondary structure and thus, pseudogenization or loss of the tRNA. If point mutations occur in the area of the anticodon, this causes an event called remolding [RBGJ10]. With a change of a base in the anticodon the tRNA possibly binds another aminco acid. Thus, the tRNA sequence and secondary structure still resembles the former tRNA but its anticodon binds another amino acid. Programs used to detect tRNAs such as tRNAscan-SE [NKE09] will include the information about the anticodon to predict the tRNA. Because secondary structure and anticodon don't fit the pattern, these sequences will not be correctly detected. We developed a method to detect such remolding events and applied this method to Drosophila species. The broader aim is to apply the method to primate species to detect possible remolding events here.

References

[BSSOAK$^+$10] Clara Bermudez-Santana, Camille Stephan-Otto Attolini, Toralf Kirsten, Jan Engelhardt, Sonja J Prohaska, Stephan Steigele, and Peter F Stadler. Genomic organization of eukaryotic tRNAs. *BMC Genomics*, 270(11), 2010.

[NKE09] Eric P. Nawrocki, Diana L. Kolbe, and Sean R. Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25:1335–1337, 2009.

[RBGJ10] Hubert H Rogers, Casey M. Bergman, and Sam Griffiths-Jones. The Evolution of tRNA Genes in Drosophila. *Genome Biology and Evolution*, 476-477(2), 2010.

Classification of reliable MetFrag candidate identifications for tandem mass spectra from selected lipid samples

Michael Witting[1], Christoph Ruttkies[2], Steffen Neumann[2], Philippe Schmitt-Kopplin[1,3]

[1] *Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany*
[2] *Leibniz Institute of Plant Biochemistry, IPB Halle, Department of Stress- and Developmental Biology, Weinberg 3, 06120 Halle, Germany*
[3] *Chair of Analytical Food Chemistry, Technische Universität München, Alte Akademie 10, D-85354 Freising-Weihenstephan, Germany*

christoph.ruttkies@ipb-halle.de

Lipid identification is a major bottleneck in high-throughput lipidomics studies. While the comparison against spectra in reference libraries is the method of choice, these libraries are far from complete.

In order to improve identification rates, the *in silico* fragmentation tool MetFrag [WSMHN10] was combined with lipid-class specific classifiers which calculate probabilities for lipid class assignments. The resulting workflow was trained and evaluated on MS/MS spectra of different commercially available lipid standard materials. This approach was then applied to MS/MS spectra of lipid extracts of the nematode *C. elegans*.

Fragments explained by MetFrag match known fragmentation pathways, e.g. neutral losses of lipid headgroups and fatty acid side chain fragments. Based on prediction models trained on standard lipid materials, high probabilities for correct annotations were achieved. The approach takes automated lipid data analysis from simple mass matching to a more reliable lipid annotation.

References

[WSMHN10] Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, and Steffen Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11(1):148, 2010.

Dynamic optimization: a powerful approach to elucidate optimality principles within metabolism

Jan Ewald, Martin Kötzing, Martin Bartl and Christoph Kaleta*
Theoretical Systems Biology, Friedrich-Schiller-Universität Jena
jan.ewald@uni-jena.de

Despite the aspiration of systems biology to be a holistic approach to understand metabolism, the dynamics are often neglected and only lowly resolved (e.g.: gene expression levels or metabolite concentrations). To uncover principles behind the dynamics of regulation we formulated time-dependent optimization problems and used dynamic optimization to obtain the optimal control of metabolism. We analyzed the time-optimal activation of metabolic pathways and synthesis of protein complexes. In this context, we observed a trade-off between sequential and simultaneous activation depending on the synthesis capacity of a cell. We could further validate this *in silico* results by comparing the operon structure of enzymes and complex subunits with the synthesis capacity of a cell. Those findings are valuable since they possibly explain not only the occurrence, but also the absence of operons in some organisms. [E+15, B+13] Additionally, we investigated the influence of metabolite toxicity on the regulation of metabolic pathways. While we observe a regulation at the first and last enzyme of a linear pathway if no toxicity is considered, the model regarding toxic intermediates reveals a change of key regulatory interactions depending on the toxicity distribution across the pathway. The optimization suggests that strong regulation of an enzyme is linked to high toxicity of products and low toxicity of substrates. The integration of a large-scale data set across various species confirms the predicted relationship between regulation and intermediate toxicity. This relation can be used to improve yield in biotechnology and also to combat pathogens by a targeted deregulation of pathways overproducing toxic intermediates.

References

[B+13] Martin Bartl et al. Dynamic optimization identifies optimal programmes for pathway regulation in prokaryotes. *Nature communications*, 4, 2013.

[E+15] Jan Ewald et al. Footprints of Optimal Protein Assembly Strategies in the Operonic Structure of Prokaryotes. *Metabolites*, 5(2):252–269, 2015.

[*] Present address: Medical Systems Biology, Christian-Albrechts-Universität Kiel

Green chemical fitness landscapes

Guillermo Restrepo and Peter Stadler
Bioinformatics Group, University of Leipzig
{guillermo,studla}@bioinf.uni-leipzig.de

It is shown how the concept of fitness landscapes, early generalised to a poset-valued framework [SF03], can be applied to the study of poset-valued chemical networks. The idea is that a chemical reaction network constitutes a configurational space, where either chemical reactions or substances involved in chemical reactions are characterised by multiple attributes, leading to an ordering of reactions based on the pairwise comparison of attributes [RV06]. We exemplify these ideas by looking for the optimal reaction route to 3-benzyl-1,3-oxazinan-2-one, a substance used in Alzheimer treatment, using green chemistry metrics for their synthetic routes. Another example looks for the "best" route to phenol, an important industrial commodity, using environmental information of the involved substances. A third example considers environmental properties of substances participating in chemical reactions where caffeine is one of the reagents. These two latter cases apply the concept of ordering substance-classes, rather than ordering substances [RB08].

References

[RB08] Guillermo Restrepo and Rainer Brüggemann. Dominance and separability in posets, their application to isoelectronic species with equal total nuclear charge. *Journal of Mathematical Chemistry*, 44:577–602, 2008.

[RV06] Rainer Brüggemann; Guillermo Restrepo and Kristina Voigt. Structure-fate relationships of organic chemicals derived from the software packages E4CHEM and WHASSE. *Journal of Chemical Information and Modeling*, 46:894–902, 2006.

[SF03] Peter F. Stadler and Christoph Flamm. Barrier trees on poset-valued lanscapes. *Genetic Programming and Evolvable Machines*, 4:7–20, 2003.

Bad Character Deletion Supertrees

Markus Fleischauer and Sebastian Böcker
Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena,
Ernst-Abbe-Platz 2, 07743 Jena, Germany
markus.fleischauer@uni-jena.de

Supertree methods combine a set of phylogenetic trees into a single supertree. Similar to supermatrix methods, these methods provide a way to reconstruct larger parts of the Tree of Life. Different from supermatrix methods, they allow us to analyze large datasets without constructing and analyzing a multiple sequence alignment for the complete dataset. Therefore, supertree methods can be used as part of divide-and-conquer meta techniques, potentially evading the computational complexity of phylogenetic inference methods such as maximum likelihood. Matrix Representation with Parsimony (MRP) is still the most widely used supertree method today, as the constructed supertrees are of comparatively high quality. Recently, the meta-method SuperFine was introduced, which combines the Strict Consensus Merger (SCM) as preprocessing with MRP and outperforms all other methods. Another recent supertree method is FlipCut [BGB13] which tries to resolve incompatibilities in the source trees by flipping '0/1'-entries in their matrix representation. FlipCut has guaranteed polynomial running time, and outperformed other polynomial-time supertree methods. Here, we introduce the Bad Character Deletion (BCD) supertree method. We remove the minimum number of clades from the source trees so that the resulting set of trees is compatible. We adapt the FlipCut heuristic for the new objective function. Furthermore, we integrate the SCM supertree into our calculations, and show how to use bootstrap values when removing bad clades. On a simulated dataset, BCD outperforms the state-of-the-art algorithms SuperFine and Matrix Representation with Parsimony. BCD supertrees simultaneously produces high-quality supertrees *and* has guaranteed polynomial running time, a combination that was previously never achieved.

References

[BGB13] Malte Brinkmeyer, Thasso Griebel, and Sebastian Böcker. FlipCut Supertrees: Towards Matrix Representation Accuracy in Polynomial Time. *Algorithmica*, 67(2):142–160, 2013.

Basal fungi phylogeny reconstructed by proteome, rRNA, RNome and mitochondrial genes

Konstantin Riege and Manja Marz
Chair of Bioinformatics, Friedrich-Schiller University Jena
konstantin.riege@uni-jena.de

The reconstructed fungal phylogeny is under continuous alteration and improvement. Especially the classification of basal fungi is a challenging task. Besides protein coding genes, currently only little is known about non-coding RNAs in fungi. We identify all known orthologous genes in between selected fungi genomes by similarity search. Furthermore we perform de novo ncRNA predictions from total RNA, using poly(A) enrichment of the mRNA, and small RNA sequencing data for *Candida albicans* and *Aspergillus fumigatus*. This transcriptome profiling improves the knowledge about their genomic landscapes. The fractions of proteins, non-coding RNAs, antisense transcripts, intronic RNAs or riboswitches are given. Originated from these analysis, we compare more than 50 fungi and calculate their phylogenetic trees with the maximum likelihood based software RAxML [Sta14] on four different gene sets: (i) Evolutionary conserved small subunit rRNA genes; (ii) Mitochondrial genomes, which are assembled by Mira [CPD+04] and annotated for homologous genes by Mitos [BDJ+13]; (iii) The entire set of non-coding RNAs, being predicted with our software pipeline Gorap; (iv) And the core proteome.

References

[BDJ$^+$13] M. Bernt, A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf, and P. F. Stadler. MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular phylogenetics and evolution*, 69(2):313–319, Nov 2013.

[CPD$^+$04] B. Chevreux, T. Pfisterer, B. Drescher, A. J. Driesel, W. E. G. Müller, T. Wetter, and Sá. Suhai. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome research*, 14(6):1147–1159, Jun 2004.

[Sta14] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9):1312–1313, May 2014.

Improving phylogenetic footprinting by neglecting evolution.

Martin Nettling[1], Hendrik Treutler[2], Jesus Cerquides[3] and Ivo Grosse[1,4]
[1] *Institute of Computer Science, Martin Luther University, Halle (Saale)*
[2] *Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle (Saale)*
[3] *Departament de Matemtica Aplicada y Anlisi, Universitat de Barcelona, Barcelona*
[4] *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig*
martin.nettling@informatik.uni-halle.de

Phylogenetic Footprinting has become increasingly attractive for motif discovery and the prediction of functional binding sites in our current next-generation-sequencing era where genomes of different species can be sequenced at an unprecedented speed. Phylogenetic relationships among the species of interest are typically represented by phylogenetic trees, and choosing an appropriate tree topology and appropriate branch lengths has been found to strongly influence the performance of motif discovery and binding-site prediction. Here, we systematically study this influence using artificial and real data. In case of artificial data, we find that the prediction performance is maximized when the tree topology and the branch lengths used for phylogenetic footprinting are equal to those used for generating the data. In case of real data, however, we find that irrespective of the studied data set the prediction performance is maximized by a tree with a star topology and infinite branch lengths, corresponding to the surprising observation that phylogenetic footprinting works best when evolution is neglected.

Whole transcriptome analysis of rare, meristematic plant cell populations using laser capture microscopy followed by strand-specific total RNA-Seq

Christoph Schuster [1, §] and Elliot M. Meyerowitz [1, 2, §]

[1] *The Sainsbury Laboratory, University of Cambridge, United Kingdom*
[2] *Division of Biology and Biological Engineering, California Institute of Technology, USA*

[§] Correspondence (christoph.schuster@slcu.cam.ac.uk
meyerow@caltech.edu)

Shoot apical meristems (SAMs) are specialized tissues at the tip of the plant, which are the source of all aboveground organs. Within the meristem, functionally distinct zones can be identified: the central zone, containing the stem cells; the organizing center, providing a niche-like microenvironment; and the peripheral zone, characterized by transit amplifying cells and the emergence of flower primordia [1]. Even though the term "meristem" exists for more than one hundred fifty years [2], the cellular differences in the gene activity of the distinct zones, as well as the transcriptional characteristics of the earliest stages of flower development, only begin to emerge [3-5]. Moreover, the machinery that controls the *de novo* formation of the stem cell system in the floral meristem is largely unknown.

To investigate the mechanisms that regulate the cellular behavior in the SAM, the early flower development, and the *de novo* formation of stem cells, we are performing a cell-type specific analysis of the distinct meristematic regions and early flower primordia in both the model plant *Arabidopsis* and in other plant species. In a first step, we have established an optimized protocol for Laser Microdissection (LMD) followed by strand-specific total RNA-Sequencing. Plant tissue after fixation and sectioning showed great RNA and tissue preservation, which is a prerequisite for successful LMD application. Libraries made from RNA dilution series of fixed and embedded tissue were sequenced and we found that most of the genes with moderate to high transcript abundance could be detected in both technical replicates. Further experiments aim to optimize the rRNA depletion and library preparation procedure. In a second step, we will use a recently described quantitative statistical method to distinguish true biological signal from technical noise to uncover the transcriptional profiles of our cell-type specific samples [6]. Inter-species comparisons of these profiles, as well as modeling the underlying regulatory networks, will ultimately lead to an understanding of the basic principles of early flower development and *de novo* formation of stem cell systems in higher plants.

References

[1] Simon Scofield and James A. H. Murray. The evolving concept of the meristem. *Plant Molecular Biology* 60: V–VII, 2006.

[2] Carl Wilhelm von Nägeli. In *Beiträge zur Wissenschaftlichen Botanik*, Erstes Heft. Leipzig, 1858.

[3] Ram Kishor Yadav, Thomas Girke, Sumana Pasala, Mingtang Xie, and Vanugopala Reddy. Gene expression map of the Arabidopsis shoot apical meristem stem cell niche. *Proc Natl Acad Sci USA* 106(12): 4941-6, 2009.

[4] Wolfgang Busch et al. Transcriptional control of a plant stem cell niche. *Developmental Cell* 18(5): 849-61, 2010.

[5] Frank Wellmer, Márcio Alves-Ferreira, Annick Dubois, José Luis Riechmann, and Elliot M. Meyerowitz. Genome-wide analysis of gene expression during early Arabidopsis flower development. *PLoS Genetics* 2(7): e117, 2006.

[6] Philip Brennecke et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* 10 (11): 1093–1095, 2013.

Optimal Block-Based Trimming for Next Generation Sequencing

Ivo Hedtke[1], Ioana Lemnian[2], Ivo Grosse[2,3], Matthias Müller-Hannemann[2]
[1] *Institute of Computer Science, Osnabrück University,* [2] *Institute of Computer Science, Martin Luther University Halle-Wittenberg,*
[3] *German Centre for Integrative Biodiversity Research*
hedtke@uos.de, {lemnian,grosse,muellerh}@informatik.uni-halle.de

Read trimming is a fundamental first step of the analysis of next generation sequencing (NGS) data. Traditionally, read trimming is performed heuristically, and algorithmic work in this area has been neglected. In [HLMG14], we have addressed this topic and have formulated three constrained optimization problems for block-based trimming, i.e., truncating the same low-quality positions at both ends for all reads and removing low-quality truncated reads. We have found that all three problems are NP-hard, have proposed three relaxed problems by omitting some of the constraints, and have developed efficient polynomial-time algorithms for them.

Here, we find that two of the three problems are even NP-hard to approximate and that we obtain only poor approximation guarantees for the third problem. Hence, we present heuristic speed-up techniques and parallelizations for the polynomial-time algorithms and study their efficacy based on twelve data sets from four species, five sequencers, and read lengths ranging from 36 to 101 bp. Scrutinizing the results of these studies, we find that (i) the omitted constraints are almost always satisfied, (ii) the optimized block-based trimming algorithms typically yield a higher number of untrimmed bases than traditional heuristics, and (iii) these results can be generalized to alternative objective functions beyond counting the number of untrimmed bases.

References

[HLMG14] I. Hedtke, I. M. Lemnian, M. Müller-Hannemann, and I. Grosse. On Optimal Read Trimming in Next Generation Sequencing and Its Complexity. In *AlCoB: Algorithms for Computational Biology, Spain, Proc.*, volume 8542 of *LNCS*, pages 83–94. Springer, 2014.

Generalized Algebraic Dynamic Programming

Christian Höner zu Siederdissen
Bioinformatics Group, Dept. of Computer Science, Universität Leipzig
choener@bioinf.uni-leipzig.de

Dynamic programming (DP) algorithms are pervasive in bioinformatics. Examples include sequence alignment, structure prediction, and phylogenetic reconstruction. While conceptually simple, actual implementations are complicated due to the need to incorporate the necessary biological details to make predictions accurate. Algebraic dynamic programming simplifies development via separation of concerns. Search space, evaluation via scoring algebras, and optimal choice are separated. This separation simplifies development considerably as algebras and choice functions are simple to implement and the search space can be described with a formal grammar.

Our *generalized* algebraic dynamic programming framework offers several advantages. (1) On the level of formal grammars we introduce their formal product. This allows simple and efficient construction of multi-tape (alignment-style) grammars. (2) gADP readily generalizes to multiple context-free grammars. This extension allows us to handle arbitrarily interleaved structures, which includes the (typically quite complicated) pseudoknotted structure prediction problem. (3) The notion of *generalized* parsing allows us to treat non-string problems, such as those involving sets, trees, or other more general structures, within one framework. (4) Ensemble-style questions, in particular those answered by the combination of an inside and an outside grammar, have simple solutions within gADP. In particular we are able to compute the outside grammar in a completely mechanical way from the given inside grammar.

Joint work with Ivo L. Hofacker, Sonja J. Prohaska, Maik Riechert, Peter F. Stadler

References

[HHS14] Höner zu Siederdissen, Hofacker, and Stadler. Product Grammars for Alignment and Folding. *IEEE TCBB*, 2014.

[HPS15] Höner zu Siederdissen, Prohaska, and Stadler. Algebraic Dynamic Programming over General Data Structures. *BMC BioInf*, 2015.

[RHS15] Riechert, Höner zu Siederdissen, and Stadler. Algebraic Dynamic Programming for Multiple Context-Free Languages. 2015.

Pervasive selection for cooperative cross-feeding in bacterial communities

Sebastian Germerodt[1], Katrin Bohl[1], Anja Lueck[1], Samay Pande[2], Anja Schroeter[1], Stefan Schuster[1], Christian Kost[2]

1) Department of Bioinformatics, Friedrich Schiller University, Jena, Germany

2) Department of Bioorganic Chemistry, Max Planck Institute for Chemical Ecology, Jena, Germany

Sebastian.Germerodt@gmail.com

Microorganisms frequently engage in metabolic interactions, in which cells release metabolites that are subsequently taken up by con- or heterospecific receiver cells. The diverse spectrum of metabolic cross-feeding interactions found in natural microbial communities involves both uni- and bidirectional interactions and the exchanged metabolite can be a metabolic by-product or be costly to produce. Despite its frequent occurrence, the conditions that favor metabolic cross-feeding as well as the population-level consequence resulting from these interactions remain poorly understood. Here we address these issues theoretically using a set of six empirically characterized genotypes that differ in their ability and propensity to produce amino acids. By systematically varying intrinsic (i.e. benefit-to-cost ratio) and extrinsic parameters (i.e. metabolite diffusion level, environmental amino acid availability) in a cellular automaton, we show that obligate metabolic cross-feeding is selected for under a broad range of conditions. Cross-feeding clusters self-organized by positive assortment in spatially structured environments. Non-producing auxotrophs are excluded from the clusters but can survive nonetheless. Strikingly, cross-feeding helped to maintain genetic diversity within populations, yet environmental supplementation of the required metabolites decoupled these obligate interactions and resulted in a loss of less competitive genotypes. Together, these results support the idea that metabolic interdependencies rapidly evolve in microbial populations.

De-Novo Transcriptome Analysis and Characterization of Juncus effusus with a Multi-Module Bioinformatics Pipeline

Martin Porsch[1,2], Stefan Michalski[3], Walter Durka[3,4], Ivo Grosse[1,4]
1: Institute of Computer Science, Martin Luther University of Halle-Wittenberg; 2: Institute of Human Genetics, Martin Luther University of Halle-Wittenberg; 3: Department Community Ecology, Helmholtz Centre for Environmental Research - UFZ; 4: German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig
martin.porsch@informatik.uni-halle.de

The natural degradation and retention of contaminants is an important part of every ecosystem in general and of wetland ecosystems in particular. The remediation potential of plants and associated microorganisms is used also as cost-effective and sustainable biotechnology. *Juncus effusus* is a representative wetland plant that degradates nitrogen. To establish *Juncus effusus* as a genetic model, we collected, characterized, and sequenced transcriptomes of several ecotypes from several locations, and we developed a modular software pipeline for the analysis of the resulting data. This pipeline involves a module for transcriptome de-novo assembly using a superposition of several transcriptome assemblers, a module for the identification of potential alleles and the computation of different allele characteristics and allele statistics, and a module for the functional annotation based on gene ontology terms and detection of homologs in NCBI's NR database. We applied this pipeline to 125 million Illumina paired-end reads and 2.8 million 454-Roche reads from one pooled sample and recovered 122,780 contigs, detected 54,175 single nucleotide polymorphisms, found 102,148 homologs in NCBI's NR database, and were able to annotate 76,970 contigs with GO-terms.

Fast and sensitive detection of differential DNA methylation

Frank Jühling, Helene Kretzmer, Stephan Bernhart, Christian Otto,
Peter Stadler, and Steve Hoffmann
University Leipzig
frank@bioinf.uni-leipzig.de

Whole genome bisulfite sequencing (WGBS) experiments allow the investigation of cytosine methylation landscapes at single CpG resolution. The correct identification of differentially methylated regions (DMRs) and, e.g., to analyse its effects on transcription, is a crucial step in epigenomic studies. Here, we present an efficient and fast algorithm to detect DMRs between groups of samples. In contrast to other methods the proposed approach does not assume any underlying distribtuions or background models and does not depend on training sets. A circular binary segmentation approach at the heart of this method is used to optimize intergroup methylation differences. In combination with a two-dimensional non-parametric test taking the methylation of multiple samples in a genomic region into account our method achieves much better results than current state-of-the-art tools. We reach true positive rates (TPR) and positive predictive values (PPV) on simulated data of typical DMRs of more than 0.99. Even on a very heterogeneous background our tool is still able to predict marginal DMRs where other tools already fail to predict most differential regions. Our approach is fast and highly memory-efficient, large sets of WGBS data of the human genome can be segmented on an usual workstation. We achieved a high user time speed-up compared to the competitors within our study while we obtain the lowest memory consumption. Additionally, the algorithm is fully parallelized, thus its computing time can be reduced to a minimum.

Posters

e!DAL - A new approach to publish your data

Daniel Arend, Jinbo Chen, Christian Colmsee,
Uwe Scholz, Matthias Lange
Leibniz Institute of Plant Genetics and
Crop Plant Research (IPK) Gatersleben
daniel.arend@ipk-gatersleben.de

Research in the area of systems biology, phenotyping, and high-throughput technologies such as next-generation sequencing are generating a massive amount of data. This data explosion enables amazing discoveries, but also arise new demands for the data management and public access. Although several international consortia and public resource centre offer services for data maintenance and management, there are shortcomings in respect of sustainability, reproducibility, integration and usability. The consequence is a high risk to loose data that was not processed due to limited resources, subjectively classified as experimental garbage or considered not to be worth to put effort into its analysis or publication respectively. Estimations suggest a data loss rate of about 80% over 20 years [GVN13].

In this contribution we discuss experiences using the e!DAL framework [ALC+14] to turn even small and mid-size institutions into public registered data centre. We developed, a handy data publication tool that provide an ad-hoc submission of research data and its registration at DataCite consortium as citable and documented scientific asset. Current data hotspot at IPK is the management of genomic and phenotypic data. We report experience from 6 month productive data publication service. So far we registered DOI's for over 50 datasets with a total volume of around 70 GB in about 21,000 files. In this period more than 80 TB was downloaded. This illustrates the public demand to get access to research data. The software, documentation and links to all productive, e!DAL powered data repositories are available at `http://edal.ipk-gatersleben.de`.

References

[ALC+14] Daniel Arend, Matthias Lange, Jinbo Chen, Christian Colmsee, Steffen Flemming, Denny Hecht, and Uwe Scholz. e!DAL - a framework to store, share and publish research data. *BMC Bioinformatics*, 15(1), 2014.

[GVN13] Elizabeth Gibney and Richard Van Noorden. Scientists losing data at a rapid rate. *Nature*, 504, 2013.

Computer-Aided Design of Small Molecules Targeting Parasitic Histone Deacetylases

J. Melesina[1], T. Heimburg[1], K. Wichapong[1], M. Schmidt[1], A. Chakrabarti[2], A. Hauser[2], C. Romier[3], R. J. Pierce[4], M. Jung[2] and W. Sippl[1]

[1]*Institute of Pharmacy, Martin-Luther University Halle-Wittenberg, Germany*

[2]*Institute of Pharmaceutical Sciences, Albert-Ludwigs University of Freiburg, Germany*

[3]*Institute of Genetics and Molecular and Cellular Biology, University of Strasbourg, France*

[4]*Institut Pasteur de Lille, University of Lille Nord de France, France*
jelena.melesina@pharmazie.uni-halle.de

Histone deacetylases (HDAC) are promising targets regulating the growth of parasitic organisms. Our aim is to develop potent and selective drug-like inhibitors of parasitic HDACs. To reach this aim, various computer-based approaches are applied, such as homology modeling, docking, and molecular dynamics simulations. Using these theoretical methods, novel inhibitors of *Schistosoma mansoni* HDAC8 were found in a virtual screening campaign [1]. Several crystal structures of the target protein have been solved later [2-3], paving a way for further structure-based optimisation of the identified inhibitors. As a result a series of HDAC inhibitors with improved activity and selectivity profile has been developed. In the current work we also focus on other parasitic pathogens, such as *Plasmodium falciparum*, *Trypanosoma cruzi* and *Leishmania major*. Homology models of their HDACs are prepared for the computer-aided design of novel selective inhibitors.

References

[1] Kannan S. et al. Discovery of inhibitors of *Schistosoma mansoni* HDAC8 by combining homology modeling, virtual screening, and in vitro validation. J Chem Inf Model. 2014, 54(10), 3005-19.

[2] Marek M et al. Structural Basis for the Inhibition of Histone Deacetylase 8 (HDAC8), a Key Epigenetic Player in the Blood Fluke *Schistosoma mansoni*. PLoS Pathog. 2013, 9(9).

[3] Stolfa DA et al. Molecular basis for the Antiparasitic activity of a Mercaptoacetamide Derivative That Inhibits Histone Deacetylase 8 (HDAC8) from the Human Pathogen *Schistosoma Mansoni*. J Mol Biol. 2014, 426 (20), 3442-53.

Modelling for activity prediction of *in silico* mutants exemplified by trisporic acid synthesis

Sabrina Ellenberger[1], Stefan Schuster[2] and Johannes Wöstemeyer[1]
[1] *Chair of General Microbiology and Microbial Genetics, Friedrich Schiller University Jena*
[2] *Department of Bioinformatics, Friedrich Schiller University Jena*
Sabrina.Ellenberger@uni-jena.de

The elucidation of regulatory mechanisms at the protein level depends on the detection of defective mutants and on functional revertants. According to the organism and peculiarities of the genetic system, this approach may be tedious or even not feasible at all.

We point out the potential of bioinformatics analysis for predicting possible regulatory mechanisms at the level of enzymatic activity. As a model, we chose an enzyme from the biosynthetic pathway towards the sex hormone trisporic acid in *Mucor*-related fungi. This system has a long tradition in biology and is well understood at the physiological level. Experimental genetic analysis is however severely hampered by special organismic properties. Biosynthesis of the hormone depends on collaboration of two complementary mating types. 4-Dihydromethyltrisporate dehydrogenase (TSP1) is a key enzyme in this pathway. Trisporic acid and its precursors are crucial for recognition between complementary mating types. In the parasitic zygomycete, *Parasitella parasitica*, these substances have an additional function. Here, they are also responsible for host-parasite interactions and the formation of infection structures. The switch between sexual and parasitic communication, which are both mediated by trisporic acid, is predominantly regulated at the protein level. Against the expectation, this fungus contains more than one gene for the TSP1 enzyme. Two of six TSP1 isoforms are able to form solid binding pockets for substrate and cosubstrate after heterodimer formation with other TSP1 isoforms. But only a single of these two enzymes can be active as homodimer. We created models for *in silico* generated mutants to investigate the decisive parts of the TSP1 protein structure that mediate activity in homodimers. Protein-protein docking was used to simulate dimerization.

Three structural elements were identified, needed for dimerization. Enzyme inactivation is achieved by closure of the active site by the long flexible loop region of the binding partner. Modification of three amino acids is sufficient to transform the inactive isoform into an active one.

Homology Modeling and Molecular Dynamics Simulations of Sirtuin 4

Zayan Alhalabi, Wolfgang Sippl
Institute of Pharmacy, Martin-Luther University Halle-Wittenberg, Germany
zayan82at@hotmail.com

SIRT4, which is localised in the mitochondria, is a member of the sirtuin family of nicotinamide adenine dinucleotide-dependent enzymes that play key roles in multiple cellular processes such as metabolism, longevity and stress response. Sirt4 regulates glutamate dehydrogenase and insulin secretion. Sirt4 also functions as a cellular lipoamidase that regulates the pyruvate dehydrogenase (PDH). Its catalytic efficiency for lipoyl and biotinyl lysine modifications is superior to its deacetylation activity (1,2,3). Recent studies reported that Sirt4 seems to have a tumor-suppressive function by restricting glutamine utilization and repressing Myc-induced Bcells lymphomagenesis (4,5), and may serve as a novel therapeutic target in colorectal cancer (6) .

Due to the absence of a crystal structure for Sirt4, a 3D model for this subtype is needed to study the possible binding mode of putative inhibitors. In the current work, a homology model of Sirt4 was generated using different templates and computational approaches. The template selection was done using HHblits - Homology detection by iterative HMM-HMM comparison technique, which identified Sirt5 and also the bacterial sirtuin Sir2-Tm as the best templates. Sequence alignment was done using program MOE, and the models were generated using program Modeller 9.11 and SwissMode. The model with the lowest value of Dope (Discrete Optimized Protein Energy) score was chosen and analysed using PROCHECK.
MD simulations were carried out on the Sirt4 models in complex with the cofactor and different substrates applying different MD simulation protocols using AMBER 12 to understand the stability and conformational changes of the modeled proteins in holo and apo forms. In addition, several MD simulations of Sirt5, for which crystal structures are available, have been performed and compared with the Sirt4 results.

References

1. Rommel A. Mathias,Elizabeth A. Rowland,Todd M. Greco,et al. (2014) Sirtuin 4 Is a Lipoamidase Regulating Pyruvate Dehydrogenase Complex Activity, Cell 159, 1615–1625.
2. Jennifer Shih and Gizem Donmez (2013) Mitochondrial Sirtuins as Therapeutic Targets for Age-Related Disorders, Genes & Cancer 4 (3-4) 91 –96.
3. Priyanka Parihar, Isha Solanki et al. (2015) Mitochondrial sirtuins: Emerging roles in metabolic regulations, energy homeostasis and diseases , Experimental Gerontology 61,130- 141.
4. VatrinetR, IommariniL, et al. (2015) Targeting respiratory complex I to prevent the Warburg effect,the International Journal Of Biochemistry & Cell Biology (63) pages 41-45.
5. Jeong SM, Xiao C, Finley LW et al. (2013) SIRT4 has tumor-suppressive activity and regulates the cellular metabolic response to DNA damage by inhibiting mitochondrial glutamate metabolism. Cancer Cell 23(4), 450–463.
6. M Miyo, H Yamamoto, M Konno,et al. (2015) Tumour-suppressive function of SIRT4 in human colorectal cancer British Journal of Cancer.

Modeling the host-pathogen interactions of the human immune system and C. albicans using game theory and dynamic optimization

Sybille Dühring, Jan Ewald, Stefan Schuster
Dept. of Bioinformatics, Friedrich-Schiller-University Jena
sybille.duehring@uni-jena.de

To understand the complex host-pathogen interactions of the human immune system with *Candida albicans*, computational systems biology approaches are very useful. *C. albicans* is one of the most important human pathogenic fungi. Alterations in the host environment can render the commensal factors of the fungus into virulence attributes once the conditions favor pathogenicity. *C. albicans* then causes infections ranging from superficial mucosal diseases and thrush in immunocompetent hosts to severe, life-threatening systemic infections in immunocompromised individuals. Those systemic infections are associated with a severe morbidity, an unacceptably high mortality and high healthcare costs. With the innate immune system as the primary line of defense against systemic fungal infections the host defense relies mainly on phagocytes, especially neutrophils and macrophages. Using mathematical modeling, particularly game theory and dynamic optimization we gain insights into the interactions of *C. albicans* and macrophages. We start by setting up a differential equation model to simulate the complex dynamics of the host-pathogen interactions and perform dynamic optimization to predict optimal regimes. We then determine pure and mixed Nash equilibria to explain why macrophages sometimes release phagocytosed Candida cells instead of killing them, a process known as nonlytic expulsion.

Structure-based Development of Novel Inhibitors for the Epigenetic Target EZH2

Abdulkarim Najjar[1], Dina Robaa[1], Wolfgang Sippl[1]

[1]Institute of Pharmacy, Martin-Luther University Halle-Wittenberg, Wolfgang Langenbeck Str 4, 06120 Halle, Germany

A.k.najjar@hotmail.com

Histone lysine methyltransferases HKMTs catalyze the transfer of a methyl group to lysine residues on histone proteins and depend on the cofactor *S*-adenosylmehtionine (SAM) as methyl donor.[1] All but one PKMTs show a conserved SET domain and can be classified according to the the methylated histone lysine residue and the level of lysine methylation (i.e. mono-, di- and trimethyl lysine).[2] EZH2 is a member of PKMTs which catalyzes Lys27 methylation on histone 3 to produce H3K27Me2/3. It has no activity on its own and needs to be in complex with other proteins to gain its ability for methylation. This complex is called Polycomb regressive complex PRC2.[3] PRC2 through its enzymatic subunit EZH2 mediates many vital activities, such as proliferation and differentiation. The association between abnormal activity of EZH2, especially overexpression, and some cancers and disorders, e.g. melanoma, lymphoma, breast, prostate and bladder cancers, has been recently substantiated.[4,5] This highlights the need for the development of EZH2 inhibitors as potential therapeutics and using EZH2 as a target for drug design.

In the present project, the focus was set on EZH2, with the ultimate goal of finding novel small molecule inhibitors of EZH2 by means of *in silico* studies. First, homology models of EZH2 were generated because the resolved X-ray structures are not adequate for use in *in silico* studies. We generated several EZH2 models by using different computational methods and employing different templates.

The models were validated before using them for virtual screening. For this purpose, a set of published inhibitors were docked into EZH2. The docking poses were analysed and the binding free energy BFE was calculated to predict the affinity of the ligands to EZH2. Enrichment studies were carried out to check the ability of the generated EZH2 homology models to enrich actives among large numbers of decoys. Finally, a structure-based virtual screen (VS) based on the generated EZH2 model was performed using several compound libraries. The VS results were analyzed to select new candidates for in vitro testing.

References:

1. Helin K, Dhanak D. *Nature*, 480 Vol. 502, 2013.
2. Wu H. et al. *PLOS ONE*, Vol. 5, Issue 1, e8570, 2010.
3. Margueron R, Reinberg D. *Nature*, 343, Vol. 469, 2011.
4. Wu H. et al. *PLOS ONE*, Vol. 8, Issue 12, e83737, 2013.
5. Tan1 J. et al. *Acta Pharmacologica Sinica*, Vol. 35, 161–174, 2014.

Machine learning in computer-aided drug design

Dat Q. Nguyen, Wolfgang Sippl
Institute of Pharmacy, University of Halle
dat.nguyen@pharmazie.uni-halle.de

Molecular docking is the most widely used method in modern computer-aided drug design, regardless of being used in virtual screen or lead optimization. Molecular docking predicts the preferred orientation of one molecule (*ligand*) to a bigger molecule (*protein*) to form a stable complex. The most complex and important step in this docking process is scoring. In this process, a large number of binding poses are computationally generated and then evaluated using a scoring function (SF), which is a mathematical or predictive model that produces a score representing binding stability of the pose. Generally, three main aspects of a SF define its goodness, they are ”docking power”, ”ranking power” and ”scoring power” [Che09]. Reportedly, conventional SFs are able to predict binding modes while mostly failed to predict binding affinities. In literature, SFs are typically classified as force-field-based, empirical, and knowledge-based. A first application of machine learning using Random Forest to predict binding affinities shows an increasing of more than 20% in term of Pearson’s correlation coefficient in a generic benchmark set with 195 protein-ligand complexes [Bal10]. Since then, a new class of SF has been intensively studied, the machine learning-based SF. More recently, this new technique was criticized showing bad performance in the ”docking power” test [Rog14]. Using Rotation Forest, an ensemble learning classifier in combination with diverse modifications by our own to guarantee diversity and accuracy, we could show that all three aspects of a SF could be greatly improved with the same benchmark set and in-house test data.

References

[Bal10] Pedro J Ballester. A machine learning approach ... *Bioinformatics (Oxford, England)*, 26(9):1169–75, May 2010.

[Che09] Tiejun Cheng. Comparative assessment of scoring functions ... *Journal of chemical information and modeling*, 49(4):1079–93, April 2009.

[Rog14] Didier Rognan. Beware of Machine Learning-Based ... *Journal of chemical information and modeling*, 54(10):2807–15, 2014.

What sequence information can reveal: The evolution of arrestins in deuterostomes

Henrike Indrischek, Peter F Stadler, and Sonja Prohaska
University Leipzig
henrike@bioinf.uni-leipzig.de

The cytosolic arrestin proteins mediate desensitization of activated G-protein coupled receptors via competitive binding of the receptor or via internalization mediated by clathrin binding motifs. As different arrestin conformations can result in specific signaling outcomes, this protein family is a possible target in drug therapeutics. The aim of the current study was to improve the existing incomplete and error-prone annotations of arrestin genes in order to reveal details about the functional evolution of these proteins. Identity and number of arrestin paralogs were determined searching vertebrate genomes or, alternatively, gene expression data with individual Hidden Markov models for each exon and paralog. Unlike standard gene prediction methods, our pipeline can detect exons situated on different scaffolds and assign them to the same gene, increasing completeness of the annotation. We uncovered the interesting duplication- and deletion history of arrestin paralogs in deuterostomes including tandem duplications, pseudogenization and retrogene formation. At the root of vertebrates, two whole genome duplications have given rise to four arrestin paralogs from a single arrestin as it is found in ciona today. An additional clathrin binding motif was gained after the duplications. The precursor of visual arrestins lost the first clathrin binding motif establishing a functional difference in arresting G-protein coupled signaling in non-visual and visual arrestins. The current work shows how an improved annotation of a multi-exon gene family can result in a detailed understanding of the link between gene architecture and functional evolution.

Phylogenetic analysis of components of the Tat (twin arginine translocation) protein transport machinery

Elisabeth Piltz[1] and Ralf Bernd Klösgen[2] and Alexander Hinneburg[1]
[1]*Institute of Computer Science and* [2]*Institute of Biology, Martin Luther University Halle-Wittenberg*
elisabeth.piltz@student.uni-halle.de

The Tat-pathway [JFM12] is able to translocate fully folded proteins into the thylakoid lumen of chloroplasts as well as across the cytoplasmic membrane into the periplasm of bacteria and archaea. This machinery consists of three proteins (TatA, TatB, TatC), whose interaction is largely unkown. Therefore, it was the aim of this work to analyse TatA and TatC in a phylogenetic way in order to gather information about the composition and cooperation of the Tat machinery.

For this purpose, a set of proteins that are homologous to TatA and TatC was collected using BLAST. Multiple sequence alignments with Clustal Ω, MUSCLE and T-Coffee were performed. Neighbor Joining algorithm [SN87] was used to visualize the phylogenetic tree. As there is no crystal structure information available for TatA, a binding site prediction model solely based on tree and conservation values was constructed.

It was possible to identify significant differences between plant and bacterial/archaeal Tat proteins. Furthermore potential binding sites and active centers could be predicted.

References

[JFM12] P Rose J Fröbel and M Müller. Twin-arginine-dependent translocation of folded proteins. *Erratum in Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 367(1599):2246, 2012.

[SN87] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

Modeling Evolutionary Shifts of Ecological Niches Based on Climatic Spaces

Chris Rothe[1], Matthias H. Hoffmann[2], Alexander Hinneburg[1]
[1] *Institute of Computer Science , Martin Luther University Halle-Wittenberg,* [2] *Institute of Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg*
chris.rothe@student.uni-halle.de

Climate niches are unknown for extinct plants or ancestors of current organisms. We collected data about living organisms that contain both climate niches and distances of suitable DNA marker sequences and propose a probabilistic latent variable models to capture correlations between the features spaces. This allows to infer possible climate niches for unobservable organisms or predict such possible niches for living organisms. In this preliminary report, we use multiple species of *Carex*, *Fagus* and *Ranunculus* and compare with previous methods [Hof05]. Correlations between genetic distances and climatic spaces of plants are modeled using a new kernelized variant of canonical correlation analysis [Hot36, DRH03]. This enables observations of linear as well as non linear relationships between genetic distances and climatic spaces, which is an additional advantage over common canonical correlation analysis. We discuss the parameterization of the model and how to interpret results in the context of evolution and geographical migration of plant species.

References

[DRH03] John Shawe-Taylor David R. Hardoon, Sandor Szedmak. Canonical correlation analysis; An overview with application to learning methods. *Technical Report*, 3(2):965–1003, 2003.

[Hof05] Matthias H. Hoffmann. Evolution of the Realized Cclimatic Niche in the Genus Arabidopsis (Brassicaceae). *Evolution*, 59(7):1425–1436, 2005.

[Hot36] Harold Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 48(3/4):321–377, 1936.

Correcting mass shifts in mass spectrometry imaging data using lock-mass-free recalibration procedure

Purva Kulkarni[1, 2]; Philipp Kynast[1, 2]; Filip Kaftan[2, 3]; Aleš Svatoš[2, 3]; Sebastian Böcker[1]

[1]Lehrstuhl für Bioinformatik, FSU, Jena, Germany; [2]MPI for Chemical Ecology, Jena, Germany; [3]Inst. of Organic Chemistry & Biochemistry, AS CR, Prague, Czech Republic

pkulkarni@ice.mpg.de

Mass Spectrometry Imaging (MSI) has seen tremendous technological advances in its application to chemical and biological samples. With every MSI experiment, it is important to maintain high mass measurement accuracy for accurate identification of the observed ions. Many times this can be compromised due to different experimental factors. Herein, we introduce a new procedure for lock-mass-free recalibration, which performs mass shift correction of spectra, one at a time, by iteratively building a crystal using spectra acquired at individual coordinate positions. Given a MSI dataset, preprocessing was performed to extract peak lists. We observe large mass deviations for identical molecules measured in different spectra. To ensure comparability between spectra and accurate mass correction, we applied a lock-mass-free recalibration procedure in three steps: First, we calculate a score for each spectrum based on similarity with its neighbouring spectra. Based on this score, an order to grow the crystal and simultaneously process the spectra is decided. Second, a consensus spectrum is generated in an iterative process which follows the crystal growth pattern and third, all spectra are corrected against this consensus spectrum. We applied our method to the data obtained from laser-assisted desorption/ionization (LDI)-TOF MSI of surface lipids on *Drosophila melanogaster* flies [Kaf14]. In this experiment, intact flies were fixed on a target plate. For this data, we observed mass deviations predominantly between ±0.3 Da and ±0.5 Da. On applying our recalibration method, we observed that mass deviations in individual mass spectra were strongly reduced. Lock-mass correction of MSI data is difficult as not all spectra contain the selected peak. Our method eliminates this need.

References

[Kaf14] Kaftan, F., Vrkoslav, V., Kynast, P., Kulkarni, P., Böcker, S., Cvačka, J., Knaden, M. and Svatoš, A., Mass spectrometry imaging of surface lipids on intact Drosophila melanogaster flies, *J. Mass Spectrom.*, 49: 223–232, 2014.

Omics for non-model organisms: the case of amphipods from Lake Baikal

Lorena Rivarola-Duarte, Christian Otto, Stephan Schreiber, Daria Bedulina, Anton Gurkov, Lena Jakob, Denis Axenov-Gribanov, Magnus Lucassen, Jörg Hackermüller, Franz Sartoris, Steve Hoffmann, Hans-Otto Pörtner, Maxim Timofeyev, Till Luckenbach and Peter F. Stadler

Interdisciplinary Centre for Bioinformatics, Universität Leipzig
Helmholtz Centre for Environmental Research, UFZ Leipzig
Baikal Research Centre and University of Irkutsk, Russia
Alfred Wegener Institute, AWI Bremerhaven
lorena@bioinf.uni-leipzig.de - studla@bioinf.uni-leipzig.de

Eulimnogammarus verrucosus and *E. cyaneus* are amphipods endemic to the unique ecosystem of Lake Baikal and serve as emerging models in particular in ecotoxicological studies. From our previous analysis on a survey sequencing of the genome of the first species [Riv14] we estimated its genome size as nearly 10 Gb. At least two thirds of its genome are non-unique DNA, and a third of the genomic DNA is composed of just five families of repetitive elements, including low-complexity sequences. We furthermore reported the assembly of the mitochondrial genome using a new, guided "crystallization" procedure, observing e.g. that it shares the gene order with the amphipod *Gammarus duebeni*. Efforts on genome size estimation from Jeffery & Gregory [JG14] showed that the c-value for *E. cyaneus* (~3.87pg) is in the range of *Homo sapiens*. Having into account the dimension and complexity of these amphipod genomes and therefore the challenging assembly, we decided in addition to sequence their transcriptomes. In our experimental design we included another palearctic amphipod species, *Gammarus lacustris*, as a mean of comparison. Exposure to heavy metals, aromatic hydrocarbons and temperature increase were the tested conditions, as possible scenarios of pollution and global warming. Illumina was used for the sequencing.

References

[Riv14] Lorena Rivarola-Duarte, Christian Otto, et al. A first glimpse at the genome of the Baikalian amphipod *Eulimnogammarus verrucosus*. *Journal of Experimental Zoology Part B*, 322(3):177-189, May 2014.

[JG14] Nicholas W. Jeffery and T. Ryan Gregory. Genome size estimates for crustaceans using Feulgen image analysis densitometry of ethanol-preserved tissues. *Cytometry Part A*, 85A:862–868, 2014.

Prediction of conserved long-range RNA-RNA interaction in full viral genomes using the example of HCV

Markus Fricke, Manja Marz
Faculty of Mathematics and Computer Science, Friedrich Schiller University Jena
markus.fricke2@uni-jena.de, manja@uni-jena.de

Long range RNA-RNA interactions (LRIs) have been reported in several RNA viruses and play important roles in the viral genome replication. LRIs are often associated with local RNA structures, such as cis-acting elements, and therewith frequently located in loop regions or internal bulges. Currently, there exists no tool which is able to predict these pseudo-knot structures in full viral genome alignments.

We present `LRIscan` a tool to predict conserved, genome-wide LRIs based on a multiple sequence alignment in only a few hours on an average computer. `LRIscan` consists of three basic steps:

(i) Identification of all possible LRIs based on a genome-wide alignment using a double sliding window approach (based on ViennaRNA Package).

(ii) The LRI-filter consists of a minimal distance of nucleotides (default: $L \geq 100$), a minimal number of interacting base-pairs (default: $d \geq 5$) called 'seed' without bulges, a conservation in $\geq 55\%$ of the considered sequences and a mean sequence complexity larger than 50%.

(iii) The final LRI score is a combination of the minimum free energy (MFE) and the structural accessibilities of the corresponding regions.

We applied our method to an alignment consisting of 106 HCV genomes in only 9.5h. We confirmed all previously known LRIs. Strikingly, we identified a conserved interaction between the apical loops of SLII and DLS. This possible initial interaction can be extended to include 62 interacting basepairs to build a potential genome circularization between the 5'UTR and 3'UTR.

With `LRIscan`, we are able to predict LRIs on multiple alignments based on hundreds of full viral RNA genomes. Our results facilitate the investigation of LRIs in the viral replication.

Preliminary comparison of differentially expressed ncRNAs in bat and human cells infected with Ebola Zaire virus

Nelly F. Mostajo Berrospi, Martin Hoelzer, Stephan Becker, Manja Marz
RNA Bioinformatics and High Throughput Analysis, Friedrich Schiller Universitaet Jena, Germany
Institut fuer Virologie, Phillips Universitaet Marburg, Germany
nelly.mostajo@gmail.com

Ebola (EBOV) is among the most lethal viruses for human. Fruit bats seem to carry EBOV without developing any symptom, however, their molecular response is not known. Non-coding RNAs (ncRNAs) are molecules which are related to several functions in the cells including viral control. The ncRNAs expression of bats during *Zaire ebolavirus* (Z-EBOV) infection could give insight into the bat responses for controlling the virus. For this study *R. aegyptiacus* (RAE) cell lines were treated with MARV, Z-EBOV, Mock (control) and studied at 3 time points. The same was done for human cell line. All the ncRNAs were searched by homology against the RFAM database [GJ03] in the RAE transcriptome. Further, the ncRNA annotation of RAE was used for cross-reference the coordinates with the mapping reads in *P. vampyrus* (closest available genome). The annotation used for humans differential expression analysis was *Ensembl* annotation (hg19). As a first overview when comparing Mock and Z-EBOV samples, there are ten differentially expressed ncRNAs in bats: U1, U2 , SNORA-73, -40, SCARNA-2, -7, -17, uc338, SRP, 7SK. These ncRNAs were also find with different patterns in humans. The ncRNAs above mentioned have different functions, including splicing and rRNA modifications, and some were previously described in viral infection. The filters used for considering these genes differentially expressed were really strict. Nevertheless, all the results are preliminary due to the use of only one cell line and the lack of repetitions.

References

[GJ03] Alex; Marshall Mhairi;Khanna Ajay; Eddy Sean R Griffiths-Jones, Sam; Bateman. Rfam: an RNA family database. *Nucleic acids research*, 31(1):439–441, 2003.

Why Respirofermentation? Explaining the Warburg effect in tumour (and other) cells by a minimal model

Philip Möller[1], Daniel Boley[2], Christoph Kaleta[3] and Stefan Schuster[1]
[1] *Friedrich Schiller University Jena,* [2] *University of Minnesota,*
[3] *Christian-Albrechts-University Kiel*
philip.moeller@uni-jena.de, stefan.schu@uni-jena.de

Tumour cells mainly rely on glycolysis leading to lactate rather than on respiration to produce ATP. This phenomenon is known as the Warburg effect (named after German biochemist Otto Warburg) and also occurs in several other cell types such as striated muscle cells, activated lymphocytes, microglia, and endothelial cells. It seems paradoxical at first sight because the ATP yield of glycolysis is much lower than that of respiration. An obvious explanation would be that glycolysis allows a higher ATP production rate, but the question arises why the organism does not re-allocate protein to the high-yield pathway of respiration. We tackle this question by a minimal model only including three combined reactions. We consider the case where the cell can allocate protein on several enzymes in a varying distribution and model this by a linear programming problem in which not only the rates but also the maximal velocities are variable. This leads to pure respiration, pure glycolysis, and respirofermentation as a mixed flux distribution, depending on side conditions and on protein costs.

Phylogenetic distribution of plant snoRNAs

Deblina Patra, Jana Hertel, Sebastian Bartschat, Ivo Grosse, and Peter Stadler
Martin-Luther-Universität Halle-Wittenberg and University of Leipzig
deblina@bioinf.uni-leipzig.de

Small nucleolar RNAs (snoRNAs) are one of the most ancient families amongst non-protein-coding RNAs whose main function is rRNA modification. Normally, snoRNAs are absent in bacteria but present in archeae hinting at the ancient origin of snoRNAs. snoRNAs are generally produced by two mechanisms; they are either directly transcribed by RNA polymerase II or produced through splicing. snoRNAs are categorized into two classes, C/D snoRNAs and H/ACA snoRNAs, which possess conserved sequence motifs. Detailed studies of these conserved sequence motifs are considered to be the first step towards deepening our understanding of the evolution of the snoRNAs. In contrast to animals, however, there is only very limited knowledge about conserved sequence motifs in plant snoRNAs.

Hence, we aim to study the phylogenetic distribution of plant snoRNAs using conserved sequence motifs. To accomplish this aim, we develop a pipeline based on the collated and phylogenetically classified group of plant species along with annotated snoRNA C/D or H/ACA boxes in each snoRNA family for each plant species. This pipeline predicts putative snoRNA candidates, snoRNA targets, characteristic snoRNA properties such as double stem loops, secondary structures, and family-wide snoRNA alignments. Finally, we perform a comprehensive study providing the complete systematic list of plant snoRNA families of phylogenetically classified groups of plants with conserved sequence motifs, their targets, and comparative sequence models.

The hidden world of non-canonical aliphatic amino acids

Maximilian Fichtner, Stefan Schuster

Department of Boinformatics, Friedrich-Schiller-University Jena
maximilian.fichtner@uni-jena.de

Amino acids are the essential building blocks of proteins and therefore living organisms. While the focus often lies on the canonical or proteinogenic amino acids there is also a large number of non-canonical amino acids to explore. Some of them are part of toxins or antibiotics in fungi or bacteria and operate, for example, like an undercover agent that sabotages on the translational level. Here we give an overview of all naturally occurring aliphatic amino acids up to a chain length of five carbons and have a closer look on each of them. Examples are (i) dehydroalanine, which is involved in lantibiotics and microcystins, (ii) 2-amino butanoic acid involved in cyclosporins, (iii) norvaline, which deteriorates the quality of several pharmaceuticals and (iv) homoleucine, being a constituent of longicatenamycines. Moreover, we outline mathematical methods for enumerating the complete list of all potential aliphatic amino acids of a given chain length. Our compilation might be interesting for numerous medical applications: the discovery of new antibiotics, design of synthetic antibiotics, improvement of protein and peptide pharmaceuticals by avoiding incorportation of non-canonical amino acids, study of toxic cyanobacteria and others.

Comparative visualization of sequence motifs with DiffLogo

Martin Nettling[1], Hendrik Treutler[2], Jan Grau[1], Jens Keilwagen[3],
Stefan Posch[1], and Ivo Grosse[1,4]
[1]*Institute of Computer Science, Martin Luther University, Halle (Saale)*
[2]*Leibniz Institute of Plant Biochemistry, Halle (Saale)*
[3]*Institute for Biosafety in Plant Biotechnology, Julius Khn-Institut (JKI), Federal Research Centre for Cultivated Plants, Quedlinburg*
[4]*German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig*
martin.nettling@informatik.uni-halle.de

Sequence motifs are widely used for the representation of functional regions of biological sequences such as transcription factor binding sites in genomic DNA, splice sites in pre-mRNAs, or phosphorylation sites of proteins. These motifs are typically visualized as sequence logos [SS]. Next generation sequencing has started to enable comparative studies of sequence motifs from different species, ecotypes, mutants, tissues, cell lines, developmental stages, or treatments in large-scale setups. In such studies, it becomes more and more important to perceive differences between sequence motifs, but these differences are often hard to detect by comparing individual sequence logos. Here, we present DiffLogo, an extensible R-package specifically designed for visualizing differences between multiple sequence motifs in an intuitive and easy-to-use manner.

References

[SS] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100.

High-throughput plant phenotyping at IPK: facilities and methods for integrated image analysis

Jean-Michel Pape, Astrid Junker, Christian Klukas, Dijun Chen and Thomas Altmann

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben

pape@ipk-gatersleben.de
junkera@ipk-gatersleben.de
klukas@ipk-gatersleben.de
chend@ipk-gatersleben.de
altmann@ipk-gatersleben.de

High-throughput (HT) plant phenotyping systems enable the quantitative analysis of a variety of plant features in a fully automated fashion. The comprehensive phenomics infrastructure at IPK comprises three LemnaTec conveyor belt-based systems for the simultaneous analysis of large numbers of individual plants of different sizes. Detailed information about the systems together with important requirements which have to be considered in the design of HT phenotyping experiments for plant cultivation in environmentally controlled conditions of model and crop plants are shown in [1]. Equipped with an integrated set of imaging sensors (RGB, NIR, static and functional chlorophyll fluorescence, IR and 3D laserscanner) these systems allow for the measurement of plant architectural, physiological and colour-related traits. The extraction of these traits from 2D images and 3D point clouds requires state of the art image processing and computer vision techniques for different classification and cluster analyses as well as post-processing pipelines for modeling and statistics [2]. A further key task is the integrated analysis of the existing multi-sensor data in order to increase the precision of phenotypic trait extraction and data interpretation. The Integrated Analysis Platform (IAP) developed for high-throughput plant image analysis at IPK represents a powerful open-source software package [3]. Based on this framework a robust leaf segmentation approach for rosette plants (*Arabidopsis thaliana* and tobacco) has been developed which can cope with different disturbances (noise, inhomogeneous background) with minimal error [4]. The poster will present an overview about IPK plant phenomics facilities and image analysis procedures applied for a variety of biological questions.

Analysis of whole-genome sequencing data for elucidating floral transition and discovery of bolting resistance genes in *Beta vulgaris*

Ioana Lemnian[1], Conny Tränkner[2], Claus Weinholdt[1], Nazgol Emrani[2], Nina Pfeiffer[2,4], Friedrich Kopisch-Obuch[2,5], Markus Schilhabel[3], Christian Jung[2], Ivo Grosse[1,6]

[1] *Institute of Computer Science, Martin Luther University Halle-Wittenberg* [2] *Plant Breeding Institute, Christian Albrechts University of Kiel* [3] *Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel* [4] *KWS Lochow GmbH, Northeim* [5] *KWS SAAT SE, Einbeck* [6] *German Centre for Integrative Biodiversity Research (iDiv) Halle-Leipzig-Jena*

lemnian@informatik.uni-halle.de

Sugar beet (*Beta vulgaris ssp. vulgaris*) is an important sucrose storing crop plant, and understanding the regulatory mechanisms of floral transition is essential for breeding winter cultivars that are bolting resistant after cold exposure. In order to identify floral transition genes underlying the recessive trait bolting resistance, we sequenced pooled DNA from F2 beets bulked for bolting versus non-bolting phenotypes after cold treatment. For the analysis of the resulting data, we developed a pipeline for detecting genomic regions that contain putative bolting resistance genes. The pipeline consists of two general modules for read mapping and SNP calling and one problem-specific module for the detection of homozygous regions in the pool of non-bolting plants compared to heterozygous regions in the pool of bolting plants. When applying this pipeline to the set of 1.1 billion paired-end reads stemming from 26 non-bolting plants and about 300 bolting plants, we find a single genomic region of about 1 Mb homozygous in the pool of non-bolting plants and heterozygous in the pool of bolting plants, which is currently scrutinized by marker analysis for the presence of putative bolting resistance genes in sugar beet.

Model-supported experimental improvement of a continuous algae production process: Minimizing medium supply via a light/dark cycle dependent dilution rate

Tobias Weise[1], Stefan Schuster[2], Michael Pfaff[1]
[1] *University of Applied Sciences Jena, Department of Medical Engineering and Biotechnology;*
[2] *Friedrich Schiller University Jena, Department of Bioinformatics*
Tobias.Weise@fh-jena.de

Economic production of algal biomass at industrial scale requires to make efficient use of natural sunlight. Algal growth and production processes are strongly influenced by light/dark cycles. In the bench-scale process studied here, on-line optical density measurements revealed an instantaneous and ongoing stagnation of algal growth during night. This is causing a fluctuation of biomass concentration under continuous operation with a constant dilution rate. This however has the disadvantage of a non-steady-state operation with respect to algal biomass. In turn, a variable dilution rate also has high potential to save culture medium. In order to model and compare these processes, the biomass balance was established using the Verhulst differential equation expanded by a term describing stagnating growth during the night, thereby obtaining a control function for a light/dark cycle dependent dilution rate. The contribution presents results from discontinuous and continuous culture experiments as starting point for the modelling as well as the theoretical introduction of a switching control function. All experiments were carried out using a strain of *Nannochloropsis salina* cultivated in a modified f/2 medium within a bench-scale tubular photobioreactor applying a 16h/8h light/dark on/off cycle. First model calculations reproduced the cyclic biomass fluctuations under a constant dilution rate. Applying, however, a dilution rate that operates only at daytime, the calculations showed a constant and higher algal biomass concentration and also considerably lower medium requirements. This has to be verified in further experiments. It is also envisaged to use the model for optimal control design.

Detecting paralog-specific gene expression associated with floral induction in the allopolyploid *Brassica napus*

Claus Weinholdt[1], Nazgol Emrani[2], Nicole Jedrusik[2], Ioana Lemnian[1], Markus Schilhabel[3], Carlos Molina[2], Christian Jung[2], Ivo Grosse[1,4]
[1] *Institute of Computer Science, Martin Luther University Halle-Wittenberg* [2] *Plant Breeding Institute, Christian Albrechts University of Kiel* [3] *Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel* [4] *German Centre for Integrative Biodiversity Research (iDiv) Halle-Leipzig-Jena*
claus.weinholdt@informatik.uni-halle.de

Oilseed rape (*Brassica napus*) is a major source of vegetable oil worldwide, and adaptation to different environments and regional climatic conditions in oilseed rape cultivars requires variation in flowering time and vernalization requirement. Hence, the identification of genetic factors that promote or inhibit flowering hold an important key for oilseed rape breeding.

Brassica napus is a modern hybrid originating from the two diploid species *Brassica rapa* and *Brassica oleracea*. Compared to the model plant *Arabidopsis thaliana*, *Brassica napus* has a redundant genome potentially containing up to six paralogs per *Arabidopsis thaliana* gene. Hence, developing an understanding of flowering time regulation in *Brassica napus* requires studying the complex interplay of highly similar paralogs along its development.

With the goal of identifying flowering time regulators of *Brassica napus* during floral induction, we developed a pipeline for the quantification of paralog-specific gene expression from RNA-seq data based on two-way ANOVA. By comparing young non-vernalized plants versus vernalized plants displaying first flowering buds, we found more than 950 paralog-specifically and differentially expressed gene groups, out of which we validated more than 30 paralogs via qRT-qPCR.

Semi-automatic interpretation of local secondary structure motifs in lncRNAs

Karolin Wiedemann[1], Jörg Hackermüller[2], Kristin Reiche[1]
[1] *Fraunhofer Instutite for Cell Therapy and Immunology (IZI), Leipzig*
[2] *Helmholtz Centre for Environmental Research (UFZ), Leipzig*
karolin.wiedemann@izi.fraunhofer.de

We, and others, found many lncRNAs to be highly specifically expressed in different tissues and disease associated pathways, proposing that lncRNAs are functional in the context of diseases [GR12, HR$^+$14]. While small ncRNAs have been grouped into families of related secondary structures, such a classification, if it exists at all, is currently not available for lncRNAs. However, evidences for purifying selection on RNA structures in mammals suggest that regulatory function of lncRNAs is defined by local secondary structure motifs. Searching for similar secondary structure motifs in a set of novel lncRNAs might be a promising approach to recognize groups of lncRNAs with related regulatory mechanisms [HC$^+$12].

Nevertheless, the interpretation of the results generated by state-of-the-art RNA secondary structure clustering tools remains difficult. We present a web application enabling semi-automatic analyses of lncRNAs with similar local secondary structure motifs. Independent from the used clustering tools, one can swiftly filter promising clusters to study them by summary statistics and graphics. Furthermore, methods for quality control and meta-analysis are included as well as integration of biological knowledge. This modularly build application is easily expandable and shareable, so that it provides a proper base for functional analysis of lncRNAs.

References

[GR12] M. Guttman and J.L. Rinn. Modular regulatory principles of large non-coding RNAs. *Nature*, 482(7385):339–46, February 2012.

[HC$^+$12] S. Heyne, F. Costa, et al. GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics*, 28(12):i224–32, 2012.

[HR$^+$14] J. Hackermüller, K. Reiche, et al. Cell cycle, oncogenic and tumor suppressor pathways regulate numerous long and macro non-protein-coding RNAs. *Genome biology*, 15(3):R48, January 2014.

`UAP` – Reproducible and Monitored Next-Generation Sequencing Data Analysis

Christoph Kämpf[1,2,3], Michael Specht[3], Sven-Holger Puppel[3], Kristin Reiche[1,3], Jörg Hackermüller[1,2]

(1) Young Investigators Group Bioinformatics and Transcriptomics, Department Proteomics, Helmholtz Centre for Environmental Research - UFZ, Leipzig
(2) Bioinformatics Group, Department of Computer Science, University Leipzig
(3) Fraunhofer Institute for Cell Therapy and Immunology, Leipzig
christoph.kaempf@ufz.de

Motivation: A major challenge in next-generation sequencing (NGS) is the reproducibility of results amongst different research groups given the same input data[NT12]. The problem arises because NGS data analysis is a multi-step process involving a plethora of bioinformatic software. The output of each step is determined by the parametrisation of the applied software. Hence, to ensure reproducibility it is essential to compactly describe and to exhaustively document the analysis.

Results: We developed `UAP`, a flexible tool for NGS data analysis, to overcome this issue and to make NGS data analysis easier to implement. The complete analysis is described via a single configuration file which can be easily exchanged between researchers. The configured analysis is implemented as a directed acyclic graph (DAG) with steps as nodes and dependencies as edges. Execution of steps propagates through the DAG, starting from source steps. The software logs during each steps execution the versions of used tools, checksums of produced files, as well as the last few kilobyte of standard out and standard error. The CPU and memory usage is monitored throughout the whole run time. This monitoring reveals bottlenecks and CPU starvations hidden in the pipeline. The software ensures the integrity of the generated data, is ready to use on clusters running Oracle Grid Engine or Simple Linux Utility for Resource Management and can be easily extended.

References

[NT12] A. Nekrutenko and J. Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672, September 2012.

P25

Searching with tandem mass spectra in molecular structure databases

Kai Dührkop and Sebastian Böcker
Chair of Bioinformatics, Friedrich-Schiller-University Jena
sebastian.boecker@uni-jena.de

The identification of metabolites, small molecules that are involved in cellular reactions, plays an important role in many areas of biology and medicine. Tandem mass spectrometry (MS/MS) is the tool of choice for the high-throughput analysis of metabolites. The automatic analysis of MS/MS data is still a challenging task. Searching mass spectra in a spectral library is a convenient strategy, but all available spectral libraries cover fewer than 20,000 compounds. However, molecular structure databases like PubChem contain more than 50 million compounds.

We present CSI (Compound Structure Identification):FingerID for searching with MS/MS data in a structure database. In a first step our method uses combinatorial optimization to compute a fragmentation tree that explains the molecular formula of the measured ion and the fragmented peaks [DB15]. In a second step we use machine learning techniques to predict a molecular fingerprint from the fragmentation tree [SDBR14]. We then search the predicted molecular fingerprint in a structure database like PubChem using a maximum a posteriori based scoring.

We have trained and evaluated our method on a dataset with 5,923 compounds using cross-validation. Searching in the PubChem database, CSI:FingerID identified 34.4 % of the compounds correctly. This is a 2.5-fold increase in correct identifications compared to existing methods for this task. We made our method available as web application.

References

[DB15] Kai Dührkop and Sebastian Böcker. Fragmentation trees reloaded. In *Proc. of Research in Computational Molecular Biology (RECOMB 2015)*, volume 9029, pages 65–79, 2015.

[SDBR14] Huibin Shen, Kai Dührkop, Sebastian Böcker, and Juho Rousu. Metabolite Identification through Multiple Kernel Learning on Fragmentation Trees. *Bioinformatics*, 30(12):i157–i164, 2014. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2014).

Combined Metabolome and Transcriptome Analysis of the Circadian Rhythm in the Cyanobacterium Synechocystis sp. Strain PCC 6803

Bertram Vogel[1,3], Sebastian Böcker[1], Manja Marz[1], Annegret Wilde[2] and Franziska Hufsky[1]
[1] *Faculty of Mathematics and Computer Science, Friedrich-Schiller-University Jena, Germany*
[2] *Institute of Biology, University of Freiburg, Germany*
[3] *Institute of Virology, Philipps-University Marburg, Germany*
bertram.vogel@uni-jena.de

Transcriptome and Metabolome analysis are two powerful techniques to gain further insights into living organisms. While there is a lot of ongoing research separately in both areas, not much is known about how they influence each other.

We want to use a combined transcriptome and metabolome approach to study the day-night cycle in the Cyanobacterium *Synechocystis sp.* Strain PCC 6803. Cyanobacteria are marine organisms and important primary producers. Metabolic engineering can be used to let cyanobacteria produce ethanol. However, the process is not very efficient at the moment.

We take six samples over the course of 24 hours and extract a transcriptome dataset by RNASeq and a metabolome dataset by Tandem-MS. In a first approach, we will try to find correlations between mRNA and metabolite abundances, possibly taking the time-resolved structure of our data into account. This will allow us to infer hypotheses in two directions. (i) Changes in the level of metabolites that are substrate/product of known enzymatic reactions might be linked to unannotated transcripts, giving a hint to their function. (ii) Known transcripts of enzyme encoding genes that correlate with unknown metabolites can help to reveal their identity.

A combined analysis of the transcriptome and metabolome is a promising approach to study the circadian rhythm of Cyanobacteria but also other external and internal factors that change the metabolism of any organism. Unknown transcripts and metabolites can possibly be annotated. This information can finally be used to improve the quality of the metabolic network of *Synechocystis sp.* Strain PCC 6803, for example to increase the ethanol production.

AnnoTALE - Identification, Annotation and Classification of Transcription Activator-Like Effectors

Annett Erkes[1], Jan Grau[1], Maik Reschke[2], Jana Streubel[2], Richard D. Morgan[3], Geoffrey G. Wilson[3], Ralf Koebnik[4], Jens Boch[2]
[1] *Institute of Computer Science, Martin Luther University Halle Wittenberg*
[2] *Department of Genetics, Martin Luther University Halle-Wittenberg*
[3] *New England Biolabs Inc., Beverly, MA 01915, USA.*
[4] *UMR 186 IRD-UM2-Cirad "Résistance des Plantes aux Bioagresseurs", BP 64501, 34394 Montpellier cedex 5, France.*
annett.erkes@informatik.uni-halle.de

Plant-pathogenic *Xanthomonas* bacteria use transcription activator-like effectors (TALEs) that bind to the promoter of plant genes and activate their transcription. *Xanthomonas* infections result in a substantial yield loss for many crop plants including rice. The binding domain of TALEs consists of tandem repeats containing two hypervariable amino acids, which are called repeat variable di-residue (RVD). Each RVD recognizes one nucleotide of its target DNA and the consecutive array of RVDs determines TALE target specificity. Here, we present AnnoTALE, an application for annotating TALEs in *Xanthomonas* genomes, for analyzing their structure and putative target genes, and for clustering TALEs by the similarity of their RVD sequences. We sequence the genome of *Xanthomonas oryzae pv. oryzae* PXO83 by PacBio sequencing and use AnnoTALE to predict all TALE genes of this strain. Building classes of TALEs from published and the newly sequenced *Xanthomonas* genomes allows us to gain new insights into TALE evolution. We compare the aligned RVDs of the class members and find that RVDs are highly conserved even on the codon level. We discover that only one to three codon pairs coding for one RVD occur in known TALEs, even though the number of theoretically possible codon pairs is substantially larger. We generate a network of possible and observed substitutions and find only one synonymous substitution between two codon pairs for RVD NN, whereas the remaining substitutions lead to a modification of the RVD. Our findings indicate that one way how TALE specificities evolve is by direct base substitutions in RVD codons.

Temperature sensitive RNAalifold - effects of using varyings temperatures on consensus structure prediction.

Stephan Bernhart, Ronny Lorenz, and Jana Hertel
Leipzig University
berni@bioinf.uni-leipzig.de

Temperature sensitive RNAalifold - effects of using varyings temperatures on consensus structure prediction. RNA secondary structure prediction is an important task in the anlaysis of RNA molecules. Evolutionary conservation of RNA structure can be used to enhance prediction methods based on thermodynamics, predicting the consensus structure of an alignment of evolutionary related RNAs. On the other hand, the thermodynamics based folding algorithms are dependent on the folding temperature, a fact that can be seen in nature and that can be reflected in the thermodynamics of single molecule prediction tools like ViennaRNA's RNAfold [LBHzS+11]. We tried to improve the performance of RNAalifold[BHW+08], ViennaRNA's commonly used consensus structure prediction tool, by adding support for different temperatures within the parameters used for the thermodynamics part of the folding. We used the CompaRNA[TLKJ13] dataset to evaluate the performance of this RNAalifold supporting multiple temperatures, applying specific (non-default) temperatures to almost 4000 sequences in about 1400 alignments. While some predictions improved because of using these temperatures, overall the effect was negligible. We conclude that increasing the quality of the thermodynamics part of consensus structure prediction is only useful in some few cases.

References

[BHW+08] Stephan H. Bernhart, Ivo L. Hofacker, Sebastian Will, Andreas R. Gruber, and Peter F. Stadler. RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

[LBHzS+11] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[TLKJ13] Puton T, Kozlowski LP, Rother KM, and Bujnicki JM. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.*, 41(7):4307–4323, 2013.

Simulation of cell and tissue dynamics using a deformable cell model

J. Neitsch[1], P. v. Liedekerke[2], S. Hoehme[1], D. Drasdo[2,1]
1) Interdisciplinary Centre for Bioinformatics, University of Leipzig, IZBI, Leipzig, Germany
2) INRIA, Le Chesnay; Cedex; France
johannes.neitsch@uni-leipzig.de

Current individual cell-based models for liver tissue [HBB$^+$10] do not resolve the cell shape in high detail. Most models use center-based forces that are determined only by cell material-constants, the radii of cells and the distance between interacting cells. For this reason, current models do not permit to investigate the precise shape of a cell and the resulting forces and tension which can be problematic in certain situations. For example, hepatocyte proliferation during liver regeneration after partial hepatectomy is known to compress micro-vessels and thereby affecting blood flow. This in turn is likely to affect liver function. Hence to better study the behavior of cells in more realistic tissue environments, it is necessary to resolve the cell shape and represent the resulting mechanical forces and tension in detail.

We have developed a novel cell model which is able to precisely resolve the cell shape based on a triangulated visco-elastic network. To calibrate the model elements and verify existing physical theories (Hertz-Theory, JKR, MD, . . .), experiments are simulated, such as deformation during pushing a single cell against a wall, pull off experiments and pressure based deformation of a single cell. To model tissue and tissue dynamics, interaction of cells and cell division is needed. The cell cycle is modeled and integrated in the deformable cell framework. Using proliferation, cell growth dynamics can be simulated and compared with agent-based spherical cells.

Interaction with other objects of complex, non-isotropic shape such as tissue structures (e.g. blood vessels), other cell models (agent-based spherical cells) or capsules are modeled within this framework. The developed deformable cell model is integrated in an existing software TiSim, which, together with TiQuant [FNJ$^+$15], constitutes CellSys, a modeling and image analysis framework.

References

[FNJ$^+$15] A. Friebel, J. Neitsch, T. Johann, S. Hammad, J.G. Hengstler, D. Drasdo, and S. Hoehme. TiQuant: Software for tissue analysis, quantification and surface reconstruction. *Bioinformatics*, 2015.

[HBB$^+$10] S. Hoehme, M. Brulport, A. Bauer, E. Bedawy, W. Schormann, R. Gebhardt, S. Zellmer, M. Schwarz, E. Bockamp, T.G. Timmel, J.G. Hengstler, and D. Drasdo. Prediction and validation of cell alignment along microvessels as order principle to restore tissue architecture in liver regeneration. *PNAS*, 2010.

Polar Plots Visualizing the Effects of Different Treatments on Gene Regulation

Yvonne Poeschl and Andreas Gogol-Döring
German Center of Integrative Biodiversity Research (iDiv)
Halle-Jena-Leipzig
Institute of Computer Science, Martin Luther University
Halle–Wittenberg
poeschl@informatik.uni-halle.de

Research in life science often focuses on the impacts of treatments on organisms. For example, the influence of drought on plants may be investigated by comparing expression levels of genes in plants grown in dry and wet conditions in order to identify genes which are significantly up or down regulated. A more complex experimental designs may involve two different treatments which could be applied either individually or together, leading to a 2x2 cross tabulation of expression levels for each gene. For example, studying the influence of drought and herbivore by comparing the expression levels of genes in plants that are grown in dry and wet conditions, without and with herbivore treatment each.

Typical aims pursued by this kind of study could be to find out which genes are significantly regulated by only one or by both treaments and whether the two treatments have in general a similar or an opposite effect on gene expression.

When interpreting the 2x2 cross tabulation as a landscape of four points in a three dimensional space, where x and y coordinates represent the two treatments and the expression levels are depicted on the z-axis, we define the direction of the gene response by the gradient which yields the steepest ascent. By fitting a plane into this landscape with minimal distances to the four points according to the least mean square principle, the gradient and the strenght of gene response can be extraced from its normal vector.

Directions and strengths of response were then visualized together in a circular scatterplot (polar plot), where each point represents one gene. From this plot it is straightforward to deduce the genes that have the largest response to a specific treatment or a combination of both treatments and the amount of their influence. Circular histograms visualizing the number of genes showing similar directions of response are available too. These plots provide an overview of the general direction of gene response in an experiment studying the influence of two treatments on gene expression.

P31

Normoxic accumulation of HIF1α is associated with glutaminolysis

Matthias Kappler[1,], Ulrike Pabst[1], Swetlana Rot[1], Helge Taubert[2], Henri Wichmann[1], Johannes Schubert[1], Matthias Bache[3], Claus Weinholdt[4], Uta-Dorothee Immel[5] ,Susanne Unverzagt[6], Ivo Grosse[4], Dirk Vordermark[3] and Alexander W. Eckert[1]

[1] Department of Oral and Maxillofacial Plastic Surgery, Martin Luther University Halle-Wittenberg, Halle(S);

[2] Department of Urology, University Hospital Erlangen, Friedrich-Alexander-University of Erlangen-Nürnberg, Erlangen;

[3] Department of Radiotherapy, Martin Luther University Halle-Wittenberg, Halle(S);

[4] Institute of Computer science, Martin Luther University Halle-Wittenberg, Halle(S);

[5] Institute of Legal Medicine Martin Luther University Halle-Wittenberg, Halle, Germany

[6] Institute of Medical Epidemiology, Biostatistics and Informatics, Martin Luther University Halle-Wittenberg, Halle(S), Germany

Abstract: The tumor prognostic factor hypoxia-inducible factor 1α (HIF1α) accumulates to a comparable extent under hypoxic and normoxic conditions, but the accumulation of HIF1α under normoxic conditions is not well understood. Here, we investigate the physiological conditions as well as the cell biological consequences of normoxic HIF1α accumulation in different tumor cell lines (e.g. MCF7, XF354, MDA-MD, SAS) by western blot analysis, microarrays and NGS. We find that the catabolism of glutamine (glutaminolysis) and the presence of toxic ammonia (a waste product of glutaminolysis) are responsible for the normoxic accumulation of HIF1α. In addition, we find that the down-regulation of the cytochrome P450 CYP1A1 is associated with this mechanism of accumulation. Third, we find that potential substrates of CYP1A1 such as aspirin are capable of destabilizing HIF1α even under hypoxic conditions, which represent an initial treatment trial. These findings suggest a key role for HIF1α in regulating the metabolism of glutamine and the level of the toxic metabolic waste product ammonia. Moreover, this findings provide new arguments for the role of ammonia in metabolic pathways associated with glutaminolysis and the Warburg effect, which opens new directions for finding therapeutic approaches that target the specific metabolism of tumor cells and of other metabolic diseases.

Modelling slime mould electrical properties by a fractional order system in the frequency domain

Andreas Mämpel, Matthias Reinecke, Michael Pfaff
University of Applied Sciences Jena,
Department of Medical Engineering and Biotechnology
andreas.maempel; johannesmatthias.reinecke@stud.fh-jena.de;
michael.pfaff@fh-jena.de

In a recent publication, Whiting et al. [Whi15] studied the electrical properties of biological wires formed by the true slime mould *Physarum polycephalum* using frequency measurements. Here, an approach is applied to model this experimental data using linear fractional order systems, i.e. systems described by linear fractional differential equations in the time domain. With respect to the Bode plot data in [Whi15], a fractional order transfer function was established in the Laplace domain that is able to adequately model the data. This can not be achieved by integer order transfer functions which represent ordinary differential equations in the time domain. More specifically, the analytical transfer function identified represents a fractional order lead-lag system with five parameters. Based on this function, the amplitude and the phase response were derived analytically in order to fit the model to the Bode plot data. Also, the analytical expressions for the Nyquist plot, i.e. the equations for the real and the imaginary part of the frequency response, were derived and the plot drawn. The inverse Laplace transform of the fractional order transfer function could however not be obtained in analytical form. Therefore, the inverse Laplace transform was calculated numerically in order to describe the system in the time domain. The approach, to our knowledge, represents the first fractional order modelling of *Physarum polycephalum* properties. This indicates that at least certain properties of the organism mirror biological fractional order systems in nature.

References

[Whi15] James G.H. Whiting, Ben P.J. de Lacy Costello, Andrew Adamatzky. Transfer function of protoplasmic tubes of *Physarum polycephalum*. *BioSystems*, 128:48–51, 2015.