German Conference on Bioinformatics 2014

Posterabstracts

Posterabstracts

1. **Daniel Arend, Jinbo Chen, Christian Colmsee, Steffen Flemming, Uwe Scholz and Matthias Lange**
   The e!DAL Java API: Sharing and Citing Research Data in Life Sciences

Abstract: Big data is one of the main challenges in life sciences. High-throughput technologies, e.g. next-generation sequencing or plant phenotyping became affordable and produce a huge amount of primary data, which is the basis for a high number of bioinformatics analysis pipelines. The data life cycle from experiments to scientific publications follows usually the schema: experiments, data analysis, interpretation, and publication of scientific paper. Beside the publication of scientific findings, it is important to keep the data investment and ensure its future processing. This implies a guarantee for a long-term preservation and preventing of data loss. Condensed and enriched with metadata, primary data would be a more valuable resource than the re-extraction from articles. It becomes essential, to change the handling and the acceptance of primary data within the scientific community. Data publications should be honoured with a high attention and reputation for data publishers. Here, we present the e!DAL Java API as a lightweight software framework for publishing and sharing of research data [ALC+14]. e!DAL was developed based on experiences coming from decades of research data management at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK). Its main features are version tracking, management of metadata, information retrieval, registration of persistent identifier, embedded HTTP(S) server for public data access, access as network file system, and a scalable storage backend. e!DAL is available as an opensource API for a local non-shared storage and remote usage to feature distributed applications.

The IPK is an approved data center in the international DataCite consortium (http://www.datacite.org) and applies e!DAL as data submission and registration system for DOIs. The e!DAL software is proven and has been deployed into the Maven Central Repository. Documentation and Software are also available at: http://edal.ipk-gatersleben.de

References

[ALC+14] Daniel Arend, Matthias Lange, Jinbo Chen, Christian Colmsee, Stef fen Flemming, Denny Hecht, and Uwe Scholz. e!DAL - a framework to store, share and publish research data. BMC Bioinformatics, 15(1):214, 2014.

2. **Markus List, Ines Block, Marlene Lemvig Pedersen, Helle Christiansen, Steffen Schmidt, Mads Thomassen, Qihua Tan, Jan Baumbach, Jan Mollenhauer**
Microarray R-based Analysis of Complex Lysate Experiments with MIRACLE

Abstract: Reverse phase protein arrays (RPPA) allow for sensitive quantification of relative protein abundance in thousands of samples in parallel. Typical challenges involved in this technology are antibody selection, sample preparation and optimization of staining conditions [SR+08]. The issue of effective sample management and data processing, however, has been widely neglected.

This motivated us to develop MIRACLE, a comprehensive and user-friendly web application covering experimental design, data processing, statistical analysis, and visualization of the results. MIRACLE bridges the gap between spotters that can create RPPAs with arbitrarily complex designs and scanners that can determine the signal of each spot, by conveniently keeping track of sample information. Raw data can be linked to pre-defined layouts and subsequently, the background corrected signal can be normalized for surface effects [NB+12], before spots belonging to one sample can be merged in a process known as quantification. To this end, several algorithms are available in R [HH+07, ZW+09]. MIRACLE integrates these methods directly through interfacing with R, which will also allow future methods to be added in a straight-forward fashion. The results are visualized and can be investigated with regards to statistical significance. In addition, experts have the possibility to export data to R for more complex analysis.

References

[HH+07] Jianhua Hu, Xuming He, et al. Non-parametric quantification of protein lysate arrays. Bioinformatics, 23:1986â94, August 2007.

[NB+ 12] E Shannon Neeley, Keith a Baggerly, et al. Surface Adjustment of Reverse Phase Protein Arrays using Positive Control Spots. Cancer Informatics, 11:77-86, January 2012.

[PC+ 01] C P Paweletz, L Charboneau, et al. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene, 20:1981-9, April 2001.

[SR+ 08] Brett Spurrier, Sundhar Ramalingam, et al. Reverse-phase protein lysate microarrays for cell signaling analysis. Nature Protocols, 3:1796-808, January 2008.

[ZW+ 09] Li Zhang, Qingyi Wei, et al. Serial dilution curve: a new method for analysis of reverse phase protein array data. Bioinformatics, 25:650-4, March 2009.

3. **Michael Dondrup, Christian Andreetta, Frank Nilsen and Inge Jonassen**
LiceBase - Building a model organism database and functional genomics tools for the sea lice research community

Abstract: Sea lice such as the salmon louse (Lepeophtheirus salmonis) are crustacean parasites of ray-finned fish, and as such pose a large challenge to worldwide aquaculture. Emerging resistance against commonly used pesticides increases the demand for novel approaches to sea lice management. Recently, the genome and transcriptome of the atlantic salmon louse have been sequenced and are being annotated.

LiceBase is a web-based system for storing, visualizing, and data-mining developed at the Sea Lice Research Centre funded and by the Research Council of Norway and ELIXIR Norway. LiceBase is based on the GMOD-tools (http://gmod.org) that also drive major model organism databases such as Fly-Base WormBase.

The establishment of Model Organism Databases (MODs) has had a large impact on the genome annotation process and MODs have become established tools for integrating research communities around them. Our aim is to establish an active global research community working jointly on the annotation of the genomes of sea lice and to collaborate in analysis of genomic data from related organisms.

Core features of our system are: a novel LIMS for annotation of RNA-interference experiments, ontologies such as the Sea Lice Developmental Stage Ontology, visualization of sea lice genomes, and related high-throughput data, and SAML-based authentication for protected or unpublished data. We are continuously collecting and integrating new resources. LiceBase developers are also seeking strong integration with the salmon genome database, and the Norwegian e-infrastructure for life science (NeLS), which are all being developed within the Norwegian ELIXIR node.

4. **Jan Krüger, Thomas Gatter, Christian Henke, Susanne Konermann, Andreas Lückner, Madis Rumming and Alexander Sczyrba**
BiBiCloud - a Cloud Computing Framework for Big Data Bioinformatics

Abstract: The need for high-throughput data analysis has grown tremendously since the introduction of next-generation sequencing (NGS) platforms. Small research labs can hardly cope with the data generated. The massive amount of data produced creates a new class of resource barriers to be overcome including limited bandwidth, storage volume and compute power. A solution to mere resource problems are private or commercially available "pay-as-you-go" clouds as a virtually unlimited and flexible resource, but the problem of making suitable software available on these resources remains to be addressed. Our BiBiCloud framework aims to support developers in setting up both server environments and tool deployment, as well as transparently integrating cloud resources. It consists of five parts: (1) BiBiServ is a web application framework that can be installed on any Unix-like system within minutes. Applications are integrated as so called BiBiApps and can be (un-)deployed during server runtime. The framework supports compute grids and has a cloud space junction for transfer of large datasets. (2) Each BiBiApp is generated from a central description that contains all information necessary to deploy a tool on the server. The BiBiApp Wizard assists the developer to annotate the tool using a pre-defined ontology. (3) The InstantBiBi module allows an easy setup of a BiBiServ instance. It automatically installs an application server, the BiBiServ framework, BiBiApp Wizard and optional extensions on a local or cloud-based machine. (4) BiBiGrid supports a flexible and scalable way to obtain cloud based compute resources at Amazon Web Services (AWS). BiBiGrid starts predefined Amazon Machine Images (AMIs) for the master and a desired number of compute nodes and configures fast storage for temporary data. (5) The BiBiS3 module provides efficient and scalable approaches to transfer data both to and from the Amazon S3 cloud storage. It features parallel "multipart" up- and download, recursive transfer of whole directories with parallelization of multiple file downloads, and simultaneous, distributed download via a cluster.

5. **Marc Bonin, Jekaterina Kokatjuhha, Sascha Johannes, Florian Heyl, Irene Ziska, Pascal Schendel, Karsten Mans, Biljana Smiljanovic, Till Sörensen and Thomas Häupl**
A Generic Online Database for Clinical Data Collection

Abstract: Background and objective: The demand for a generic structural collection and optimal sharing of clinical data is increasing due to the large amount of data and constantly changing requirements. The system should offer the flexibility in adding new parameters as well as criteria, to perform new analysis without extra programming effort. It should also provide the collection, processing and sharing of the data in agreements according to data privacy regulations and at the same time be accessible through the intranet/internet. Materials and Methods: Programming was based on a web framework in Ruby on Rails. SQLite was used as database system. The complete software package is running on a Linux or Mac OSX, which acts as a normal server. Results: With the new software, users can generate in a standard browser tables by defining the names of rows and columns and enter a type of data corresponding to each field of the table. The possible data types include function, images, files, string, numbers, dates, enumeration with the possibility of adding default values to each of them. The interface is divided into administrative work for creating or deleting the tables, rows and columns, specifying advanced view parameters of the tables, and into data collection for entering new patients, visits as well as data values. The database is currently applied for collecting information in the multi-center biomarker research project ArthoMark. To provide data policy, different levels of rights were generated for reading, writing and sharing data as well as administrating the structure of tables. Any changes applied to the data and tables are tracked in the log file. For example, a data structure for clinical information from patients with arthritis was established. Conclusion: Any kind of information can be stored in the new database. Every new parameter can be added without programming knowledge. The database is simultaneously accessible from the Internet. Thus the database enables to collect data in the clinics, to share these with scientists, to perform biobanking or sample tracking. The database is currently part of the BMBF funded national research network ArthoMark and the EU funded network BTCure.

6. **Balarabe R. Mohammed, Craig S. Wilding, Phillip J. Collier and Yusuf Y. Deeni**
Bioinformatic Analysis of Regulatory Elements within the Promoter Region of the Cytochrome P450 gene, CYP6M2 in Anopheles gambiae

Abstract: Cytochrome P450s including CYP6M2 have been demonstrated to be involved in the metabolism of insecticides, typically through up regulation in resistant individuals. The difference in gene expression levels seen in insecticide resistant mosquitoes may result from sequence differences in the 5' upstream region of CYP6M2 including those in the promoter elements. Understanding the complex mechanisms regulating P450 expression, including that of CYP6M2 in insecticide resistant Anopheles gambiae remains a great challenge. In this study, extensive bioinformatics resources were used to predict regulatory elements and cross-species comparison in the cis-acting elements within CYP6M2 (896 bp) and CYP6G1 (896 bp) known to be up regulated by the orthologs of Nuclear factor erythroid 2- related factor 2(Nrf2)/Kelch -like ECH-associated protein 1(Keap 1) and Aryl hydrocarbon receptor (AhR)/ Aryl hydrocarbon nuclear translocator (ARNT) in Drosophila melanogaster. Searches were also made for the cis- acting elements within a 879 bp region up stream of CYP6M2 hypothesised to contain the promoter for this gene in both the Tiassale multiple insecticide-resistant and Kisumu susceptible strains of An. gambiae. Results revealed the presence of Nrf2/Keap 1 and AhR/ARNT as putative transcription factor binding sites (TFBS) within the CYP6M2 promoters. Further, we identified the orthologs of those transcription factors which bind to these elements (Cap 'n' collar isoform C (CnCC) / (Keap 1) & Spineless (ss)/ Tango in Drosophila melanogaster) as AGAP010295/AGAP000748 & AGAP005300/ AGAP003645 in Anopheles gambiae. These data suggest the presence of putative AGAP010295/AGAP009748 and AGAP005300/ AGAP003645 binding sites in the promoter of Anopheles gambiae CYP6M2, which may potentially be associated with the up regulation of CYP6M2 involved in insecticide resistance. These if established have implications in the control of malaria.

7. **Florian Halbritter and Simon Tomlinson**
Identification of Functionally Related Genomic Elements by Similarity Clustering of ChIP-seq Profiles

Abstract: Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has been used extensively to profile the interactions of proteins (e.g. transcription factors) with DNA on a global level. To ease access to and promote reuse of publicly available datasets, we have previously built up an integrated database of thousands of ChIP-seq datasets, providing an accessible compendium of regulatory data (http://www.geneprof.org [HKT14, HVT12]).

In order to further utilize these data to derive novel insight, we now summarize the genomic profiles mined from this database to define regulatory signatures of DNA-binding proteins and histone modifications pertaining to all genes and transcriptional elements of the genome. With these regulatory signatures at hand, we demonstrate how a simple hierarchical clustering-based approach can be used to identify functionally related genes. Seeding this method with known inputs helps to discover novel candidates with a role in biological pathways of interest or in disease.

We illustrate the approach by applying it to data from mouse embryonic stem cells. Importantly, many known key components of the core transcriptional circuitry of these cells are successfully recovered, substantiating the validity of the method. Dozens of less well known candidates offer promising targets for future research.

References

[HKT14] F. Halbritter, A. I. Kousa, and S. R. Tomlinson. GeneProf data: a resource of curated, integrated and reusable high-throughput genomics experiments. Nucleic Acids Res., 42(Database issue):D851-858, Jan 2014.

[HVT12] F. Halbritter, H. J. Vaidya, and S. R. Tomlinson. GeneProf: analysis of high-throughput sequencing experiments. Nat. Methods, 9(1):7-8, Jan 2012.

8. **Minou Nowrousian, Stefanie Traeger, Stefan Gesing and Daniel Schindler**
Comparative genomics and transcriptomics of filamentous fungi

Abstract: Filamentous fungi are eukaryotic microorganisms with a woldwide distribution in nearly all ecosystems. They have a huge impact as saprobes, pathogens or symbionts, and are important model organisms in cell and molecular biology as well as biotechnology. Fungi were among the first organisms to have their genomes sequenced with next-generation sequencing techniques, and have been widely used as models to establish sequencing-based methods to analyze biological questions. We have sequenced and assembled the genomes of the filamentous fungi *Sordaria macrospora* and *Pyronema confluens* with a combination of Illumina/Solexa and Roche/454 sequences. We are now using comparative genomics and transcriptomics to study the evolution of multicellular development in fungi in a number of approaches. (i) Comparative transcriptomics allows the identification of conserved patterns of gene expression, which can be used for the identification of candidate genes for downstream analyses, because conservation of expression is a strong predictor of functional significance. We have used comparative transcriptomics of the filamentous fungi Sordaria macrospora, Fusarium graminearum, and Neurospora crassa to identify developmentally regulated genes whose functions were verified experimentally. (ii) Furthermore, we could show in a comparison of RNA-seq data from seven fungal species that global gene expression levels in fungi do not follow a common mode of distribution, in contrast to what was shown for animals. (iii) RNA-seq data were also used to identify antisense transcripts in non-strand-specific sequencing data. (iv) Comparative genomics in combination with transcriptomics was used to show that orphan genes in the filamentous fungus Pyronema confluens are preferentially upregulated during sexual development, consistent with a hypothesis of rapid evolution of sex-associated genes.

9. **Andreas Bremges, Tanja Woyke and Alex Sczyrba**
Metagenomic proxy assemblies of single cell genomes

Abstract: Over 99% of the microbial species observed in nature cannot be grown in pure culture, making it impossible to study them using classical genomic methods. Metagenomics and single cell genomics are two approaches to study the microbial dark matter.

Metagenomics can obtain genome sequences from uncultivated microbes through direct sequencing of environmental DNA. Each genome's metagenomic coverage is constant and depends only on its abundance. A complementary approach to sequencing DNA of a whole microbial community is single cell genomics. Prior to sequencing of a single cell, its DNA needs to be amplified. This usually is done by multiple displacement amplification (MDA), introducing a tremendous coverage bias. Poorly amplified regions result in extremely low sequencing coverage or physical sequencing gaps. These parts of the genome cannot be reconstructed in the subsequent assembly step, and therefore genomic information is lost.

Today, it is feasible to generate both, single amplified genomes and a corresponding metagenome, from the same environmental sample. We developed a fast, k-mer based recruitment method to sensitively identify metagenomic "proxy" reads representing the single cell of interest, using the raw single cell sequencing reads as recruitment seeds. By assembling metagenomic proxy reads instead of the single cell reads, we circumvent most challenges of single cell assembly, such as the aforementioned coverage bias and chimeric MDA products. In a final step, the original single cell reads are used for quality assessment of the proxy assembly.

On real and simulated data we show that, with sufficient metagenomic coverage, assembling metagenomic proxy reads instead of single cell reads significantly improves assembly contiguity while maintaining the original accuracy. By applying our method iteratively, we span physical sequencing gaps and are able to recover genomic regions that otherwise would have been lost. However, careful contamination screening is needed.

10. **Henrike Indrischek, Sonja Prohaska and Peter Stadler**
Reconstructing the gene phylogeny of arrestins: annotation of multi-exon genes

Abstract: The cytosolic arrestin proteins mediate the desensitization of G-protein coupled receptors and activate effector molecules within intracellular signaling cascades [LGP10]. Since detailed phylogenetic information is incomplete for arrestins, this study aims at obtaining annotation of arrestin homologs in major animal orders with the goal to detect the familiy's last common ancestor. Each of the four paralogs of human arrestin is encoded by 15 or 16 relatively short exons interleaved with intronic regions of lengths up to 60.000 nt. As a consequence, automatic methods frequently fail to correctly annotate arrestin genes. Starting off with manually improved annotations of arrestins in mammals and sauropsids, we built exon- and paralog-specific profile hidden Markov Models on protein level. We developed a pipeline semi-automatically annotating homologous protein sequences in three steps. First, we score the species' translated genome with the models, second, hits are assigned to paralogs based on bit scores and synteny, third, the protein sequence is annotated considering splice sites. Applying the pipeline, we could identify four arrestin paralogs in Sarcopterygii. Within the mammalian superorder Afrotheria one paralog appears to be lost or severely truncated suggesting at least a partial complementation of the missing function by retained paralogs. Due to the teleost specific genome duplication, we find six to seven arrestin paralogs in Danio rerio and Takifugu rubripes. The emergence of the four paralogs could be traced back to known whole genome duplications in early vertebrate evolution. Hence, the vertebrate arrestins form a distinct group separated from basal deuterostome arrestins.

References

[LGP10] L. M. Luttrell and D. Gesty-Palmer. Beyond desensitization: physiological relevance of arrestin-dependent signaling. Pharmacol. Rev., 62(2):305-330, Jun 2010.

11. **Corinna Ernst and Sven Rahmann**
A Density-based Approach for the Identification of Differentially Methylated Regions

Abstract: Hyper- or hypomethylation of multiple CpGs in close spatial proximity, i.e. CpG islands, determines a broad range of biological mechanisms, e.g. cell differentiation, carcinogenesis and inflammation [Boc12]. However, recent tools for identifying differentially methylated regions (DMRs) from bisulfite sequencing data are rare and limited to special use cases. We present a novel approach for DMR detection which is based on differential methylation densities (DMDs). Given a genomic region containing several differentially methylated CpGs, we define the region's DMD as the sum over all degrees of differential methylation of incorporated CpGs divided by the region's length. These values constitute a suitable criterion for delineating DMRs, as they depend on (1) the individual degrees of differential methylation per CpG, (2) the distances between individual CpGs and (3) the concordance of methylation tendencies (i.e., hypomethylation or hypermethylation). In order to identify DMRs we use an adapted version of Chung's linear-time algorithm for the maximum density segment problem [CL05] to search for regions including local maxima, respectively minima, of differential methylation densities. Our approach is adapted to the analysis of whole-genome as well as reduced representation bisulfite sequencing data and does not rely on the availability of multiple samples per case and control group.

References

[Boc12] C. Bock. Analysing and interpreting DNA methylation data. Nature Reviews Genetics, 13:705-719, 2012. [CL05] K.-M. Chung and H.-I. Lu. An Optimal Algorithm for the Maximum- Density Segment Problem. SIAM Journal on Computing, 34(2), 2005.

12. **Sabrina Ellenberger, Lisa Siegmund and Johannes Wöstemeyer**
A new task for HGT Calculator: Exploring relations between lifestyle and the frequency of horizontal gene transfer in Protozoa

Abstract: Protozoa are a widespread group of organisms with high diversity and the ability to colonize many different habitats. Even within a genus, different modes of living are frequently observed. We can differentiate between free-living, symbiont-harbouring, parasitic, and pathogenic protozoa.

We investigate the establishment of artificial, experimentally amenable endosymbiotic relationships between the ciliate Tetrahymena pyriformis and food bacteria. The acquisition of genes by horizontal gene transfer is an important factor in such endosymbiotic relationships. We are interested in possible differences in probability and frequency of horizontal gene transfer from bacteria into protozoa caused by living in special relationships.

We assume that free-living protozoa taking up bacteria as food particles, have a higher frequency of horizontal gene transfer than parasitic ones, which often have a reduced genome and low selective pressure on additional genes because of depending on the host and because of defect complementation by the host. During digestion, sometimes food bacteria may escape from food vacuoles and subsequently transfer part of their DNA to the protozoon nucleus.

We developed a new application for specific detection of horizontal gene transfer, the HGT Calculator, which makes it possible to check, whether a gene transfer event known from one protozoon has also occurred in a closely related protozoan group or not. For this, we do not need large data sets (in worst case complete genomes) or a species tree as for other detection tools, for example, RIATA-HGT or TREX. We work on the sequences of interest and a small reference set from fungi, plant, metazoa, and prokaryotes for these genes.

We included three different taxonomic groups (amoeba, alveolata, and trypanosomatids) and different lifestyles (free-living, non-parasitic, parasitic, pathogenic, symbiont-harbouring).

13. **Ulrike Löber, Ljerka Lah, Joachim Selbig and Stefanie Hartmann**
Genome assembly and annotation of the phytopathogenic fungus Ophiostoma bicolor

Abstract: The fungus Ophiostoma bicolor is a conifer pathogen which uses bark beetles as a vector to infect trees. The genomes of two North American bark beetle associated fungi have already been sequenced and annotated (Gros- mannia clavigera and Ophiostoma piceae [HWL+13]). Genomic profiles of these species help explain their tolerance of host tree defense chemicals like monoterpens [LHBB13]. We de novo assembled the O. bicolor genome with the Velvet assembler and annotated the genome using the MAKER2 pipeline and a reference database of fungal nucleotide and protein sequences and O. bicolor RNA-Seq data. We examined assembly quality by comparing the full read set and a read set normalized using BBnorm. Previously, it was shown that increasing read coverage leads to a stagnation of assembly statistics [HBBH11]. Sequencing errors accumulate when genomes are of ”too high” coverage, which leads to mis-assemblies resulting in fragmentation of contigs. In this work, we show that using non-normalized data with very high coverage could have additional negative effects on the genome assembly.

References

[HBBH11] Sajeet Haridas, Colette Breuill, Joerg Bohlmann, and Tom Hsiang. A biologist’s guide to de novo genome assembly using next- generation sequence data: A test with fungal genomes. Journal of microbiological methods, 86(3):368-375, September 2011.

[HWL+ 13] Sajeet Haridas, Ye Wang, Lynette Lim, et al. The genome and transcriptome of the pine saprophyte Ophiostoma piceae, and a comparison with the bark beetle-associated pine pathogen Grosmannia clavigera. BMC Genomics, 14(1):373, June 2013.

[LHBB13] Ljerka Lah, Sajeet Haridas, Joerg Bohlmann, and Colette Breuil. The cytochromes P450 of Grosmannia clavigera: Genome organization, phylogeny, and expression in response to pine host chemicals. Fungal Genetics and Biology, 50:72-81, January 2013.

14. **Veronika Dubinkina, Alexander Tyakht and Dmitry Alexeev**
Study of applicability limits for k-mer methods in metagenomic data analysis

Abstract: Sequence comparison plays a key role in various sequencing data analyses. With the development of next-generation sequencing (NGS) technologies, a large amount of short read data has been generated. In recent years among the methods for genomic data analysis, the researcher's interest has been attracted to alignment-free methods for comparing sequences based on work with k-mers (oligonucleotides of length k, also called l-tuples or n-grams). This approach is highly effective for exploratory analysis (e.g. taxonomic identification and clustering simple metagenomic data) of large data sets, since it requires a relatively small number of calculations in comparison with the other methods and does not depend on comparison with the reference sequence set. However, for such complex metagenomes as gut microbiota, these methods are not yet sufficiently developed and therefore the purpose of this study was to investigate the limits of applicability of the method and the k-mers space for metagenomes. The database consisted of 339 human gut metagenomic samples sequenced in large-scale projects[1, 2, 3]. Descriptive analysis and comparison of the k-mer with reference-based bacterial compositions were performed at various values of k. Using simulated and experimental data, the contribution of various factors to the similarity of the k-mer and bacterial composition was assessed. Recommendations were developed for the application of k-mer method in data analysis of human gut metagenome.

References

[1] Qin J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464:59-65, 2010.

[2] Qin J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490:55-60, 2012.

[3] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature, 486:207-214, 2012.

15. **Ben C Stöver and Kai F Müller**
LibrAlign - A Java library with powerful GUI components for multiple sequence alignment and attached data

Abstract: Several applications currently developed in our group and by cooperators deal with multiple sequence alignments (MSA) or associated raw and meta data, and allow the user to view and edit these in a graphical user interface. Instead of implementing independent solutions for these different tasks, we decided to develop LibrAlign, a Java library providing reusable and extendable GUI components for MSA data.

LibrAlign is fully interoperable with BioJava but its flexible architecture also allows any other storage model for the (alignment) data. A Swing and a SWT version is provided for each component, so LibrAlign can be used in all Java GUI applications, including projects based on the Eclipse Rich Client Platform (RCP) or Bioclipse.

The main visual component of LibrAlign (AlignmentArea) displays a MSA and allows efficient manual editing. Additionally, so-called data areas can be attached to each sequence or to the alignment as a whole and display associated data (e.g. trace files, positions of repeat patterns, comments, consensus sequences) directly in the alignment. Custom data areas can easily be created by application developers to deal with new types of data.

Among the projects based on LibrAlign currently under development are (i) the Taxonomic Editor of the EDIT platform (which uses the Eclipse RCP), (ii) the alignment editor PhyDE, (iii) AlignmentComparator (which visualizes differences between alternative MSAs of the same dataset), and (iv) HIR-Finder (which locates microstructural mutations like tandem repeats possibly associated with hairpins).

Software and poster download: http://bioinfweb.info/LibrAlign

16. **Linda Sundermann, Sebastian Jünemann and Jens Stoye**
ChimP - Chimera Prediction with the Jumping Alignment Algorithm

Abstract: Amplicon sequencing is often used to analyze the composition of a bacterial community, using e.g. the 16S rRNA gene. Prior to sequencing, the DNA is amplified via PCR where chimeras can arise. These are artificial, single-stranded DNA sequences formed from parts of other DNA sequences, called the parents of a chimera. When not removed from the data, chimeras can have a falsifying influence on the research study. We present ChimP (Chimera Prediction), a method that identifies chimeras and their parents in a set of next-generation amplicon sequencing reads with a de novo approach. The algorithm identifies a chimera by aligning a query sequence to a multiple sequence alignment (MSA) of similar sequences with the Jumping Alignment algorithm [SRS02]. If the alignment of the query 'jumps' between the rows of the MSA, and a score threshold is reached, the query is marked as chimeric and its parents are determined. To find sequences similar to the query for the MSA beforehand, a modified version of the SWIFT filtering algorithm [RSM06] is used. We evaluated ChimP on different data sets against UCHIME [EHC+11], that is currently the most accurate chimera detection tool for next-generation amplicon sequencing data. We show that ChimP has a higher sensitivity than UCHIME.

References

[EHC+11] Robert C Edgar, Brian J Haas, Jose C Clemente, Christopher Quince, and Rob Knight. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27(16):2194-2200, 2011.

[RSM06] Kim R Rasmussen, Jens Stoye, and Eugene W Myers. Efficient q- gram filters for finding all Î-matches over a given length. Journal of Computational Biology, 13(2):296-308, 2006.

[SRS02] Rainer Spang, Marc Rehmsmeier, and Jens Stoye. A novel approach to remote homology detection: jumping alignments. Journal of Computational Biology, 9(5):747-60, 2002.

17. **Jinbo Chen, Christian Colmsee, Maria Esch, Matthias Klapperstück, Eva Grafahrend-Belau, Matthias Lange and Uwe Scholz**
LAILAPS: an integrative information retrieval platform for plant genomic resources

Abstract: Here we present the information retrieval (IR) system LAILAPS [LSB+10] and its application for plant genomic information retrieval in the frame of the European transPLANT consortium http://www.transplantdb.eu.

A central aspect is to extract knowledge from integrated genomic resources to assist a forward genetic research, such as the selection of candidate genes within a quantitative trait loci (QTL) to identify a particular region of the genome that is associated with the trait of interest.

Information retrieval becomes an indispensable method to extract knowledge from a continuously increasing number of genomic databases. Although the advent of well cited IR-systems, there is still a lack of platforms for plant genomic that support relevance models to select the best performing candidate genes that are associated with a specific trait.

This motivated the development of LAILAPS, a comprehensive IR platform for plant genomes. LAILAPS comprise a corpus of around 65 million indexed documents over 13 major life science databases and around 80 million links to genomics resources. This enables a comprehensive keyword based query system. In order to rank and select trait relevant records from thousands of hits an artificial neural network is applied. It was trained on a set of 400 expert curated and relevance ranked gene annotations for 20 different trait queries. The relevance prediction scores 11 most discriminating features that where determined by domain experts and evaluated for QTL candidate gene prediction.

LAILAPS is available at http://lailaps.ipk-gatersleben.de

Acknowledgments: This work was made in the frame of the transPLANT project and is funded by the European Commission within its 7th Framework Programme, contract number 283496.

References

[LSB+ 10] Matthias Lange, Karl Spies, Joachim Bargsten, Gregor Haberhauer, Matthias Klapperstück, Michael Leps, Christian Weinel, Röbbe Wünschiers, Mandy Weißbach, Jens Stein, and Uwe Scholz. The LAILAPS Search Engine: Relevance Ranking in Life Science Databases. Journal of Integrative Bioinformatics, 7(2):e110, 2010.

18. **Mohammad Tagi Hasanzada and Laila Hassanzada**
Epigenetic alignment in HDAC acid amine families

Abstract: Epigenetic regulation of acid amine is dynamic and reversible process that establishes normal cellular phenotypes but also contributes to human diseases. Epigenetic regulation involves hierarchical covalent modification of DNA and proteins that package DNA, such as histones. Here, we study histones deacytylases (HDACs) proteins (enzymes). HDACs use acetyl-coA as a cofactor and catalyse reverse an acetyle group of lysine residues to restore their positive charge and stabilize the local chromatin architecture.

Using Multiple Sequence Alignment and Phylogenetic construction methods, a hypothetical evolutionary relationship was generated between togethers the neighbor alignment HDACs Protein sequences and may thay nearest HDACs sequences together there drugs effect near them? We selected 29 aa of HDAC sequences. HDAC sequences are collected from NCBI databeses. We used bioinformatics techniques, sequence analyses and phylogenetic tree algorithms. This evolutionary relationship can help us to better understand similarity drugs mechanism, between nearest HDACs sequences. And these may help us to better understand drugs side effect mechanism.

19. **Marcel Bargull, Kada Benadjemia, Benjamin Kramer, David Losch, Jens Quedenfeld, Sven Schrinner, Jan Stricker, Dominik Köppl, Dominik Kopczynski, Henning Timm, Johannes Fischer and Sven Rahmann**
Variant Tolerant Read Mapping with Locality Sensitive Hashing

Abstract: The rapid development of genomic sequencing technologies in the past decades has outgrown the advances of computing power and therefore requires efficient read mapping algorithms [LBB12]. Read mappers align sequenced reads to a reference genome, where a set of reads aligned to the same position can hint at possible mutations. Even though many fast read mappers have been published in the recent years, most of them do not consider common variants of the reference genome. Variant tolerance highly increases accuracy of read mappers when aligning reads against a species' pangenome. We have developed a new read mapper for variant tolerant alignment by usage of hash based filtering in combination with an alignment algorithm based on dynamic programming. In the first step, we use locality sensitive hashing (LSH), initially designed for finding similarities in documents, for candidate filtering [AI06]. We treat reads and windows of the reference genome as documents, which are compared by LSH. As a result, we obtain an approximative mapping to the reference regions. This leads to a dramatic reduction of the reference length and therefore semi-global alignment becomes reasonable. The aligner handles variants like SNPs, insertions and deletions and decides which variants lead to the best alignment. New genetic variants and gene mutations can be found by observing the mismatches from the alignments.

References

[AI06] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on, pages 459-468. IEEE, 2006.

[LBB12] Po-Ru Loh, Michael Baym, and Bonnie Berger. Compressive genomics. Nature biotechnology, 30(7):627-630, 2012.

20. **Dirk Willrodt, Florian Markowsky and Stefan Kurtz**
Unified Methods for Blast Searches on Compressed Sequence Databases

Abstract: Most genomes currently sequenced are highly similar to those already collected. This holds, for example, for individual human genomes, tumor genomes, or for different strains of the same bacteria. Similarly, collec- tions of protein sequences are full of redundancies. As a consequence, the amount of new DNA and protein sequences is growing much more slowly compared to all available sequences. While there are many approaches ex- ploiting theses redundancies to reduce the storage capacity of the sequence data, few approaches exist that exploit the redundancies for speeding up sequence analysis. The recent papers by Loh et. al [LBB12] and Daniels et al. [DGP+13] have shown how to speed up Blast searches as follows: Split the sequence data into unique sequences on one side and redundant sequences encoded relative to the unique sequences on the other side. This serves two purposes: firstly, the efficient encoding of the redundant data reduces the space needed for those sequences. Secondly, an algorithm like Blast can be applied in a two stage fashion, one run on the unique data set and another one on the results of this coarse search and all corresponding sequences from the redundant compressed sequences. While [LBB12, DGP+13] have developed two different methods and tools, we implemented a more unified method and tool Condenser, which is alphabet-independent and can handle DNA as well as protein sequences. Moreover, it employs a optimized encoding of redundant data to improve its compression and decompression efficiency. We evaluated Condenser for different datasets of varying redundancy. Condenser achieves similar compression rates and runtime as the tools of [LBB12, DGP+13] but with a lower memory peak so that it can handle larger datasets.

References

[DGP+13] N. M. Daniels, A. Gallant, J. Peng, L. J. Cowen, M. Baym, and B. Berger. Compressive genomics for protein databases. Bioinformatics, 29(13):i283-i290, Jul 2013.

[LBB12] P.-R. Loh, M. Baym, and B. Berger. Compressive genomics. Nat Biotech, 30(7):627-630, 2012.

21. **Guillaume Holley, Roland Wittler and Jens Stoye**
Bloom Filter Trie - a data structure for pan-genome storage

Abstract: High Throughput Sequencing technologies have become fast and cheap in the past years. As a result, large-scale projects started to sequence tens to several thousands of genomes per species. The concept of pan-genome has therefore emerged. It is composed of two parts: The core genome, which represents the pool of all the genes shared by all the strains of the species, and the dispensable genome, which represents genes not shared by all strains. Graphs exhibit good properties for representing a pan-genome, as they avoid storing redundant information and can represent genomic variants. Unfortunately, existing tools using this representation are slow or memory consuming. We present here a new data structure for storing pan-genomes: The Bloom Filter Trie. This data structure allows to store a pan-genome as an annotated de-Bruijn graph in a memory efficient way. The insertion of new genomes does not degrade the performance of the structure.

22. **Peter Großmann**
The Bacterial Supercoiling Level - Modeling and Implications

Abstract: DNA supercoiling has been known as a high-level global regulator of gene expression for quite some time, yet how gene expression responds to changes in the supercoiling level is still unpredictable, except for few cases. My work is an approach to this gap from the systems level perspective, integrating different omics-datasets to break the genome down into supercoiling domains to infer local properties of the DNA that are determined by the local supercoiling level. To this end, an analytical model of supercoiling dynamics is presented, yielding a relationship of domain length, supercoiling level and RNA chain elongation rate. The model is complemented by the inference of the RNA chain elongation rate from genome-scale data.

23. **Markus Lux, Barbara Hammer and Alexander Sczyrba**
An Efficient Pipeline for Automatic Grouping in Single-Cell Sequencing and Metagenomics

Abstract: New technologies allow for ultra fast sequencing of probes in emerging areas such as single-cell sequencing or metagenomics. A crucial step in the analysis of these probes is the detection of clusters representing the involved species – this allows for an initial binning in case of metagenomics or contamination detection in single-cell sequencing. For this task, manual screening is often tedious and therefore there is a strong need for automated methods to boost the analysis process. Vectorial representations of the sequences such as k-mer frequencies often act as first preprocessing step based on which machine learning techniques such as a low dimensional linear projection towards the main principle components can be used. In this contribution, we investigate the suitability of a novel, non-linear dimensionality reduction technique, Barnes-Hut SNE [vdM13], which puts a particular focus on the inference of cluster structures and which can be run in almost linear time. Sophisticated clustering methods can profit from this first step and are able to automatically determine the number of clusters in this context. For this task we implemented a pipeline: Non-linear, fast dimension reduction methods are paired with multiple clustering algorithms and confidence measures based on clustering stability [VL10]. We compare different possible choices and report the results in benchmark examples. Further, we point out possible extensions and improvements such as the inclusion of side information.

References

[vdM13] Laurens van der Maaten. arXiv:1301.3342, 2013.

[VL10] Barnes-Hut-SNE. arXiv preprint Ulrike Von Luxburg. Clustering Stability. Now Publishers Inc, 2010.

24. **Jonas Ibn-Salem, Sebastian Köhler, Michael I Love, Ho-Ryun Chung, Ni Huang, Matthew E Hurles, Melissa Haendel, Nicole L Washington, Damian Smedley, Chris J Mungall, Suzanna E Lewis, Claus-Eric Ott, Sebastian Bauer, Paul Schofield, Stefan Mundlos, Malte Spielmann and Peter N Robinson**
Deletions of chromosomal regulatory boundaries are associated with congenital disease

Abstract: Recent data from genome-wide chromosome confirmation capture analysis indicate that the human genome is divided into conserved megabase-sized self-interacting regions called topological domains. These topological domains form the regulatory backbone of the genome and are separated by regulatory boundary elements or barriers. Copy-number variations can potentially alter the topological domain architecture by deleting or duplicating the barriers and thereby allowing enhancers from neighboring domains to ectopically activate genes causing misexpression and disease, a mutational mechanism that has recently been termed enhancer adoption.

We use the Human Phenotype Ontology database to relate the phenotypes of 922 deletion cases recorded in the DECIPHER database to monogenic diseases associated with genes in or adjacent to the deletions. We identify combinations of tissue-specific enhancers and genes adjacent to the deletion and associated with phenotypes in corresponding tissue, whereby the phenotype matched that observed in the deletion. We compare this computationally with a gene-dosage pathomechanism that attempts to explain the deletion phenotype based on haploinsufficiency of genes located within the deletions. Up to 11.8% of the deletions could be best explained by enhancer adoption or a combination of enhancer adoption and gene-dosage effects.

Our results suggest that enhancer adoption caused by deletions of regulatory boundaries may contribute to a substantial minority of copy number variation phenotypes and should thus be taken into account for their medical interpretation.

25. **Guokun Zhang, Sebastian Schaaf and Ulrich Mansmann**
ConDetec - Detecting and removing impurities in NGS data sets

Abstract: Biological samples taken for performing NGS data generation are expected to be of a 'pure' origin. In a medical context, samples should contain human cells only. In reality, due to impure nucleic acid preparations or infected samples containing pathogens, contamination which contains sequences from other species can occur unnoticed. The most common contamination sources are from microbial sequences which have greatest impact on the downstream NGS analysis: not only wasting time but also causing erroneous conclusions [Res].

Tools which can remove such contamination sequences are therefore required. Traditional methods (e.g. VecScreen [NCB]) relying on NCBI BLAST are limited to small data size and therefore not suitable for NGS data. Here we created a bioinformatics tool called "ConDetec", based on the fact that NGS data sets can be analyzed for sequences which indicate for a microbial origin. Due to the great phylogenetic distance between humans and microbes, sequences which are unique to microbes should not appear in databases of annotated human sequences. Public pathogen databases with human context, e.g. PATRIC [ea14], were used to offer microbial reference sequences. Alignment tools designed for NGS data, e.g. BWA [HR10], were employed to map raw NGS reads to the human genome reference and the microbial database. After identification of sequences of microbial origin, those can be separated from the human sequences, elevating the raw data quality. This tool is generic and also applicable for metagenomic data analysis if the focus is pathogen discovery.

References

[ea14] Wattam et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucl Acids Res, 42(D1):581-591, 2014.

[HR10] Li H. and Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, 26(5):589-595, 2010.

[NCB] NCBI. VecScreen: Screen a Sequence for Vector Contamination.
http://www.ncbi.nlm.nih.gov/tools/vecscreen/

[Res] NCBI Resources. Contamination in Sequence Databases.
http://www.ncbi.nlm.nih.gov/tools/vecscreen/contam/

26. **Paulo Pinto and Erich Bornberg-Bauer**
Moulding genotype-phenotype maps, moulding evolution

Abstract: The rate at which new phenotypes are acquired is a function of the stability of the wild-type phenotype and of its robustness to mutations. In evolution, molecular species, like organisms, must strike a balance between robustness and evolvability, between resisting and allowing change in their genome. For if robustness implies a degree of safety against deleterious mutations and replicating errors, it can also prevent the emergence of mutations with a beneficial phenotypic expression [Wag08].

We construct complete genotype-phenotype mappings for a lattice protein model (HP model) [CWBBC02] and short RNAs [FW12] (18-24 nucleotides) allow the full investigation of the genotype (sequence) space. All sequences are folded and the native conformations found are denominated the corresponding phenotype. Genotype-phenotype maps form networks where nodes represent sequences with an associated fold (phenotype) and edges are mutations. All nodes connected by neutral mutations (leaving the phenotype unchanged) form a neutral network, the acquisition of new phenotype can be translated to a mutation leading out of such a neutral network.

Beyond the ground state conformation (the native fold) of a HP protein or RNA molecule, there are higher energy conformations which can also be found albeit with a lower probability [BBC99]. We find that roughly 74% of the all RNA sequences have a stable fold. In the case of HP proteins, we find that sequences that fold into degenerate states are very common ($\sim$97.9% of all sequences) and they are mostly unstable. We mould RNA and HP model neutral networks by progressively removing the sequences with the less stable folds from the genotype-phenotype mapping. This procedure reduces the size of the neutral networks and the number of neighbouring phenotypes but, for the most part, leaves intact the average robustness — the probability of making a neutral mutation. More important is the reduction on the average fragility — the probability that a mutation is non-neutral — due to the reduction of the total number of neighbours surrounding a genotype.

The escape time from a neutral network is identical in the RNA and HP models for identical levels of fragility. However, other quantities influence evolvability — measured as the number of generations it takes for a population evolving on the genotypes forming a neutral network to reach a random neighbouring phenotype. These include the size of a neutral network and the number of different phenotypes neighbouring it, the number of mutational paths available for a genotype (the number of neighbours on the graph) and the number of phenotypes that can be reached from it. For networks containing only and all sequences with stable native folds, we measure quite distinct fragility values with correspondingly different levels of evolvability. By moulding the genotype-phenotype mappings, we create a set of neutral networks which display identical fragility but quite distinct evolvability. Using the topological measures just mentioned we model the evolvability and find that it reproduces the results obtained by direct simulation of an evolving Moran population.

References [BBC99] Erich Bornberg-Bauer and Hue Sun Chan. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. Proceedings of the National Academy of Sciences, 96(19):10689-10694, September 1999.

[CWBBC02] Yan Cui, Wing Hung Wong, Erich Bornberg-Bauer, and Hue Sun Chan. Recombinatoric exploration of novel folded structures: a heteropolymerbased model of protein evolutionary landscapes. Proceedings of the National Academy of Sciences of the United States of America, 99(2):809-814, January 2002.

[FW12] Evandro Ferrada and Andreas Wagner. A comparison of genotype- phenotype maps for RNA and proteins. Biophysical journal, 102(8):1916-1925, April 2012.

[Wag08] Andreas Wagner. Robustness and evolvability: a paradox resolved. Proceedings. Biological sciences / The Royal Society, 275(1630):91-100, January 2008.

27. **Nina Luhmann, Cedric Chauve, Jens Stoye and Roland Wittler**
Scaffolding of Ancient Contigs and Ancestral Reconstruction in a Phylogenetic Framework

Abstract: The knowledge about the structure of ancient genomes can shed light on the evolutionary processes underlying the development of extant genomes. Recent progress in sequencing also ancient DNA, so called paleogenomes, allows the integration of this sequencing data in genome evolution analysis. However, the assembly of ancient genomes is fragmented because of DNA degradation over time, resulting in a challenging scaffolding step. We address the issue of genome fragmentation in the assembly within a phylogenetic framework while also improving the reconstruction of other ancient genomes in the phylogeny. The tree is augmented at one point of the phylogeny with the fragmented assembly represented as an assembly graph.

Our approach is to compare the ancient data with extant related genomes within a phylogeny and to reconstruct genomes that minimize the single-cut-or-join rearrangement distance along the tree. In contrast to most rearrangement distances, Feijão and Meidanis showed that the minimization of the SCJ distance over the tree can be computed in polynomial time for conflict-free data.

Although the ancient DNA data is not conflict-free, we can still include these information in the reconstruction while ensuring consistent results. Using the Hartigan algorithm [Har73], we can generalize the result of [FM11] towards multifurcating trees and also include edge lengths to avoid a sparse reconstruction.

[FM11] Pedro Feijão and Joao Meidanis. SCJ: a breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(5):1318-1329, 2011.

[Har73] John A Hartigan. Minimum mutation fits to a given tree. Biometrics, pages 53-65, 1973.

28. **Johannes Köster and Sven Rahmann**
An algebra of single nucleotide variants

Abstract: The detection of mutations from Next-Generation Sequencing data is usually achieved with a call and filter approach: First, a variant caller infers variants in the sequenced sample compared to a given reference genome. The significance of a variant is often provided as a posterior probability. Second, these variants are filtered, e.g. against those of a different tissue (e.g. filtering the variants of a tumor sample against one or more normal samples) or other individuals (e.g. filtering the variants of a child against those of its ancestors). One of the key problems of call and filter approaches is to control the false discovery rate, since the posterior probabilities do not reflect the applied filters.

We present ALPACA, an algebraic single nucleotide variant caller that incorporates a flexible filtering mechanism directly into the variant calling. ALPACA implements a novel algebraic procedure that allows to estimate the posterior probability for observing a variant that behaves like specified in a given query formula. The query formula can be used to flexibly model filtering scenarios. Since the posterior probability reflects the filtering, it can be used to effectively control the false discovery rate. ALPACA preprocesses each sample separately into an index and makes use of massive parallelism with OpenCL on either the CPU or the GPU, such that different filter scenarios and thresholds can be explored within seconds.

29. **Jan Grau, Jens Boch and Stefan Posch**
Genome-wide TALEN off-target prediction

Abstract: Transcription activator-like effector nucleases (TALENs) have become an accepted tool for targeted mutagenesis. The DNA binding domain of TALENs originates from transcription activator-like effectors (TALEs) of plant-pathogenic Xanthomonas bacteria. The DNA-binding domain of TALEs and thus TALENs is composed of conserved amino acid repeats containing repeat-variable diresidues (RVDs) that determine DNA binding specificity. In TALENs, this DNA-binding domain is fused with a Fok1 endonuclease domain, where TALEN dimers specifically cut the DNA double strand. Although TALENs cut DNA highly specific, undesired off-targets remain an important issue, since these may cause severe side effects due to off- target mutagenesis. We developed a tool for the genome-wide prediction of TALEN off-targets, named TALENoffer [GBP13]. TALENoffer is based on the same statistical model as TALgetter [GWR+13], a program that successfully predicts TALE targets in host promoteromes. In TALENoffer, this model is employed to score TALEN monomers. Monomer scores are combined into a dimer score that compensates for differing monomer lengths and that allows for partial compensation between monomers for target site mismatches. In benchmark studies, TALENoffer yields a competitive performance compared to alternative approaches. TALENoffer features strategies for runtime optimization, which allow to scan complete genomes for TALEN off-target sites within a few minutes on a standard PC. We make TALENoffer available as a web-application at http://galaxy. informatik.uni-halle.de and as a command line program at http://jstacs.de/index.php/TALENoffer.

References

[GBP13] Jan Grau, Jens Boch, and Stefan Posch. TALENoffer: genome-wide TALEN off-target prediction. Bioinformatics, 29(22):2931-2932, 2013.

[GWR+13] Jan Grau, Annett Wolf, Maik Reschke, Ulla Bonas, Stefan Posch, and Jens Boch. Computational Predictions Provide Insights into the Biology of TAL Effector Target Sites. PLoS Comput Biol, 9(3):e1002962, 03 2013.

30. **Johannes Köster and Sven Rahmann**
The ParallEl AligNment UTility (PEANUT) for GPU-based read mapping

Abstract: PEANUT (ParallEl AligNment UTility) is a highly parallel GPU-based read mapper with several distinguishing features:

(1) a novel q-gram index (called the q-group index) with small memory footprint built on-the-fly over the reads. The index can be accessed and built in a massively parallel way using prefix sums and population counts. To the best of our knowledge, this is the first feasible GPU-only implementation of a q-gram index.

(2) the ability to efficiently output both the best and all hits of a read. PEANUT outperforms other state of the art CPU and GPU based read mappers in both fields. For all hits of a read, up to 10x speedups are achieved on ordinary gaming GPUs.

(3) several options to compute and output alignment quality.

A preprint describing PEANUT in detail is available[KR14].

References

[KR14] Johannes Köster and Sven Rahmann. Massively parallel read mapping on GPUs with PEANUT. Preprint: arXiv:1403.1706, 2014.

31. **Hendrik Schäfer, Tim Schäfer, Joerg Ackermann, Claudia Döring, Sylvia Hartmann, Martin-Leo Hansmann and Ina Koch.**
Hodgkin Lymphoma – From Image Analysis to Cell Graphs

Abstract: Hodgkin lymphoma (HL) is a type of B cell lymphoma which usually arises from germinal center B–cells. To diagnose the disease and identify the specific subtype, biopsies are taken and immunostained. The slides are then scanned to produce high-resolution digital whole slide images (WSI). Thus, large and growing databases of these images are currently emerging. We are interested in the automatic and systematic analysis of the WSIs to determine morphological and immunohistochemical features of HL cells and their spatial distribution in lymph node tissue. In this work, we first describe an imaging pipeline that can be used to identify stained CD30 cells in the large images. Then we then define special geometric graphs based on the positions and morphological properties of the stained cells. We analyze and describe both local and global properties of these cell graphs.

32. **Sascha Daniel Krauß, Dennis Petersen, Daniel Niedieker, Erik Freier, Samir F. El-Mashtoly, Klaus Gerwert and Axel Mosig**
Label-free Identification of Organelles through Colocalization of Raman and Fluorescence Microscopic Images

Abstract: A major promise of Raman microscopy is the label-free detailed recognition of cellular and subcellular structures. For this purpose, a key step to ”annotate” these subcellular components in Raman spectral images is to identify colocalization patterns between Raman and fluorescence microscopic images. While existing approaches rely on fluorescence labeling, we propose a combination of a colocalization scheme with subsequent training of a supervised classifier that allows label-free resolution of cellular compartments. This colocalization scheme unveils statistically significant overlapping regions by identifying correlation between the fluorescence color channels and clusters from unsupervised machine learning methods like hierarchical cluster analysis (HCA). The scheme is used as a pre-selection to gather appropriate spectra as training data, which are used in the second part as training data to establish a supervised Random Forest classifier to automatically identify organelles, e.g. lipid droplets and nucleus. We validate these classifier results by overlaying them with fluorescence labelings of different cellular compartments, indicating that many components may indeed be identified label-free in the spectral image.

33. **Thomas Temme, Martin Schmuck, Axel Mosig and Ellen Fritsche**
Novel computational approaches for High Content Image Analyses (HCA) of organoid neurosphere cultures in vitro.

Abstract: Developmental neurotoxicity testing (DNT) according to current guidelines is highly time- and cost-intensive. Therefore, combining in-vitro methods facilitating the DNT testing processes with quantitative fluorescence image analysis is indispensable. The 3D-"Neurosphere model" represents an adequate system for DNT testing since it is able to mimic specific endpoints of brain development, in particular proliferation, migration, differentiation into neurons, astrocytes and oligodendrocytes as well as apoptosis. Making such complex organoid systems applicable for substance screening, medium throughput evaluation of these endpoints is mandatory. Therefore, we propose novel microscopic image analysis algorithms to evaluate endpoints like neuronal quantification, neurite outgrowth and neurospherespecific endpoints like "radial cell migration" and "distance-dependent density distributions" with high accuracy and precision. Our algorithms automatically exclude areas of non-allocable cell densities present in the sphere core and the area nearby. For neuronal identification, the overlap of the cell nucleus channel and neuron channel is analyzed in combination with morphological features that reflect essential properties of neuron growth. Our neuronal identification is reaching an average detection power (DP) of 80-85% (versus manual evaluation) and a false positive ratio (FP) of 10-15% improving the results (DP: 50% and FP: 40%) of the commercially available Neuronal Profiling bioapplication (Thermo Scientific), which was initially designed for pure neuronal cultures. Identified coordinates of cell nuclei and neuron cell bodies are further utilized to evaluate the neurospherespecific endpoints. Coordinates of cell nuclei were used to assess the outer rim of the migration area defined by the furthest migrated cells to obtain the average radial migration distance. In combination with the coordinates of neuron cell bodies, detected by our novel algorithms, the neuronal density distribution can be calculated. In conclusion, HCA of neurospheres is a promising technique for medium throughput screening to be used in safety and efficacy testing in the future.

34. **Jan Kölling, Karin Gorzolka, Karsten Niehaus and Tim W. Nattkemper**
Spatio-temporal analysis of metabolite profiles during barley germination

Abstract: Mass spectrometry imaging generates a series of localized mass spectra from discrete positions on a tissue or thin-film, thereby providing comprehensive information on molecular composition and spatial distribution in a single experiment. This allows for an untargeted and simultaneous measurement of a wide variety of molecules.

To study a biological process over time and space, multiple MSI measurements can be conducted and combined into a time series. The sample preparation and imaging process are destructive for the sample, therefore each time point needs to be represented by a different specimen. As a result the sample morphology is different for each time point and spatial distribution can not be directly compared. Inter and intra sample variance of the signal intensity is also an issue.

After standard MSI preprocessing steps we use H^2SOM clustering of manually selected mass signals for all time points with our previously developed tool WHIDE [KLA+12]. The resulting hierarchy of cluster prototypes is further analyzed to identify spatio-temporal patterns across the time series.

The approach was applied to localize metabolites during the barley germination process which is of high scientific and agricultural interest.

[KLA+12] Jan Kölling, Daniel Langenkämper, Sylvie Abouna, Michael Khan, and Tim W Nattkemper. WHIDEa web tool for visual data mining colocation patterns in multivariate bioimages. Bioinformatics, 28(8):1143-1150, 2012.

35. **Chen Yang and Axel Mosig**
An Algorithm for the Registration of Vibrational Microspectroscopic Images in Histopathological Stains

Abstract: As established registration algorithms established for matching images of other modalities seem to fail in solving this superimposition problem, overlays have commonly been obtained by manual registration using image editing software, eventually with the help of marking landmarks positions in the sample.

We present a robust and efficient registration algorithm that reliably and in a fully automated fashion registers infrared microscopically measured regions within whole slide H&E staining images. We demonstrate that our approach works reliably on different tissue types (colon, bladder and lung) and is robust against the inherent heterogeneity of the H&E staining process.

Our approach requires several major generalizations, adaptations, and improvements of existing method. Our approach fully automates an important step that is commonly conducted manually in many spectral histopathology workflows, and we believe it will help simplifying such workflows and make them more efficient. Furthermore, our approach tasks involving other combinations of microscopy platforms.

36. **Brijesh Singh Yadav, Pavan Kumar Yadav and Ajay Kumar**
Structural and functional characterization of TIMP-3 protein in mammary tumor of Canis lupus familiaris

Abstract: In dogs (Canis lupus familiaris), mammary tumors are the second most common tumors (after skin tumors) and the most common in female dogs. TIMP-3 (Tissue Inhibitor of Metlloproteinases-3) is a matrix associated endogenous inhibitor of MMPs (Matrix Metalloproteinases). These two protein complexes are consideration to be involved in progression of tumor, and provide structural support for tumor growth. TIMP-3 gene from dog mammary tumor tissue was amplified, sequenced (Gene bank accession number is JX144398.1) and characterized. TIMP-3 protein has a cDNA sequence of 567 bp codifying a protein with 188 amino acid residues, corresponding to 21690Da and Grand average of hydropathicity (GRAVY) is 0.424. The 3D (three dimensional) structure of TIMP-3 is predicted based on homology modeling. The predicted structure is verified using appropriate software's and refined by molecular dynamic simulations. This study will facilitate in understanding the structural and functional basis of TIMP-3 in canine mammary tumors.

37. **Kutub Ashraf**
An Immunoinformatics approach for designing epitope based vaccine strategy against S protein of mysterious new Middle East Respiratory Syndrome Coronavirus (MERS-CoV

Abstract: In 2012, a deadly virus namely Middle East Respiratory Syndrome Coronavirus (MERS-CoV) has emerged from the Arabian Peninsula and it is striking fear in the hearts of public health officials throughout the world. Recent studies find that, similar to SARS-CoV, the spike (S) protein of MERS-CoV also plays important roles in receptor binding and viral entry that affects viral host range. As the major protein causing virus infection, S protein can be an ideal target for both vaccines and MERS-CoV entry inhibitors. Hence, analyzing the properties of MERS-CoV S protein is a high research priority [1,2]. As there is no effective drug available, novel approaches regarding epitope prediction for vaccine development were performed in this study. In this study, we identified several immunodominant sites on the S protein by by immunoinformatics tools. Epitopes or peptide fragments as nonamers of these antigenic S proteins were analyzed by according to their proteasomal cleavage sites, TAP scores and IC50<250 nM, the predictions were scrutinized. Furthermore, the epitope sequences were examined by in silico docking simulation with different specific HLA receptors. This study suggests that the S protein is highly immunogenic and induces protection against MERS -CoV challenge and that neutralizing antibodies alone may be able to suppress virus proliferation, further advocating the rationale that vaccines against MERS-CoV can be evolved based on the S protein, which can provide high population coverage.

References

1. de Groot RJ, Baker SC, Baric RS, Brown CS, Drosten C, Enjuanes L, Fouchier RA, Galiano M, Gorbalenya AE, Memish ZA, Perlman S, Poon LL, Snijder EJ, Stephens GM, Woo PC, Zaki AM, Zambon M, Ziebuhr J. Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. J Virol. 2013;87(14):7790-2.

2. Memish ZA, Zumla AI, Al-Hakeem RF, Al-Rabeeah AA, Stephens GM. Family cluster of Middle East respiratory syndrome coronavirus infections. N Engl J Med. 2013;368(26):2487-94

38. **Dominik Kopczynski and Sven Rahmann**
An Online Peak Extraction Algorithm for Ion Mobility Spectrometry Data

Abstract: Ion mobility spectrometry (IMS), coupled with multi-capillary columns (MCCs), has been gaining importance for biotechnological and medical applications [WLF+09] because of its ability to measure volatile organic compounds at extremely low concentrations in the air or exhaled breath at ambient pressure and temperature. Ongoing miniaturization of the devices creates the need for reliable data analysis on-the-fly in small embedded low-power devices.

We present an automated online peak extraction method for MCC/IMS spectra. Each individual spectrum is processed as it arrives, removing the need to store a whole measurement of several thousand spectra before starting the analysis, as is currently the state of the art. Thus the analysis device can be an inexpensive low-power system such as the Raspberry Pi.

The key idea is to extract one-dimensional peak models (with four parameters) from each spectrum and then merge these into peak chains and finally two-dimensional peak models. We describe the different algorithmic steps and evaluate the online method against state-of-the-art offline peak extraction methods [BVB08] that use a whole measurement.

[BVB08]B. Bödeker, W. Vautz, and J. I. Baumbach. Peak finding and referencing in MCC/IMS-data. International Journal for Ion Mobility Spectrometry, 11(1):83-87, 2008.

[WLF+ 09] M. Westhoff, P. Litterst, L. Freitag, W. Urfer, S. Bader, and J.I. Baumbach. Ion mobility spectrometry for the detection of volatile organic compounds in exhaled breath of lung cancer patients. Thorax, 64:744-748, 2009.

39. **Sabrina Ellenberger, Anke Burmester and Johannes Wöstemeyer**
The mtDNA of the mycoparasitic fusion parasite Parasitella parasitica: Sequence and comparative analysis

Abstract: Parasitella parasitica represents a parasexual system that is studied, because the infection process implies the transfer of Parasitella genes to the host. This parasexual system has been studied for nuclear genes. No information is available for the fate of mitochondrial genes after invading the host's cytoplasm. As a prerequisite for studying the spreading of mitochondrial information among Mucor-like fungi, we sequenced the mtDNA of P. parasitica.

The mitochondrial genome of P. parasitica is a single circular molecule with a length of 83,361 bp and a GC content of 30%. It is larger than other zygomycetous mitochondrial genomes: Rhizopus oryzae (54,178 bp), Mortierella verticillata (58,745 bp), or Smittium culisetae (58,654 bp). Fifteen protein coding genes were identified, as well as 26 tRNA genes, and the genes for the large and small rRNAs. All fifteen mitochondrial protein coding genes are conserved between P. parasitica and R. oryzae, with amino acid pairwise sequence identities between 80% and 100%. The nad6 has the lowest similarity followed by nad1. Other genes showed more than 85% amino acid identity between P. parasitica and R. oryzae. The nad4L gene was the most conserved, showing 100% sequence identity between P. parasitica and R. oryzae. The genome structure of P. parasitica, R. oryzae, M. verticillata, and S. culisetae is not conserved, although some genes are clustered in P. parasitica and R. oryzae: [atp9, cox3, cob, cox2, atp6], [nad3, nad2, cox1], [nad4L, nad5], [nad4, atp8], and [nad1, nad6]. Seven genes have twice as many exons in P. parasitica as in R. oryzae. The highest number of exons can be found in cox1 with eight exons. Another feature of this mitochondrial genome is the high number of putative homing endonucleases in introns.

40. **Shailendra Gupta, Ulf Schmitz, Xin Lai, Julio Vera and Olaf Wolkenhauer**
Cooperative miRNA regulation in anticancer drug resistance - A computational approach

Abstract: MicroRNAs (miRNAs) are small 22nt long functional RNA molecules, that regulate gene expression at the post-transcriptional level. We found that they can cooperatively regulate mutual targets [1, 2]. We have previously observed interplay of miRNAs with E2F1, a versatile transcription factor involved in many critical cellular processes, which can mediate resistance to anti-cancer drugs. We present here a computational workflow that integrates bioinformatics, structural and kinetic modelling approaches to simulate the effects of cooperative E2F1 regulation by miRNA pairs and possible consequences on drug-resistant tumour cells. From the five possible miRNA pairs that cooperate for E2F1 regulation identified through our computational pipeline, it is suggested that the miRNAs miR-205 and miR-342-3p can effectively repress E2F1 in a cooperative manner than other miRNA pairs. We performed kinetic modelling, followed by three-dimensional molecular docking and simulations, which support the hypothesis of a RNA triplex formation with two miRNAs and the E2F1 mRNA. Furthermore, we expanded our previously proposed kinetic model of the E2F1-p73/DNp73-miR-205 network [3] by including miR-205 and miR-342-3p cooperative miRNA pair to identify possible mechanisms that promote genotoxic drug-induced apoptosis. Our results suggest that therapies that combine genotoxic drug administration with silencing of either miR-342-3p or miR-205, can enhance the drug-induced apoptosis of tumor cells with extreme overexpression of E2F1.

References

1. Lai X, Schmitz U, Gupta SK, Bhattacharya A, Kunz M, Wolkenhauer O, Vera J: Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs. Nucleic Acids Res 2012.

2. Schmitz U, Lai X, Winter F, Wolkenhauer O, Vera J, Gupta SK: Cooperative gene regulation by microRNA pairs and their identification using a computational workflow. Nucleic Acids Res 2014.

3. Vera J, Schmitz U, Lai X, Engelmann D, Khan FM, Wolkenhauer O, Putzer BM: Kinetic Modeling-Based Detection of Genetic Signatures that Provide Chemoresistance via the E2F1-p73/DNp73-miR-205 Network. Cancer Res 2013.

41. **Jessica Kaufmann, Zhiqin Huang, Andrius Serva, Barbara Burwinkel, Peter Sinn, Andreas Schneeweiss, Peter Lichter, Marc Zapatka and Niels Grabe**
Identification of patient-individual, immunologically addressable combinatorial cancer targets through a database driven RNA-seq bioinformatics pipeline

Abstract: In the last decade, since the advent of cetuximab, clinical cancer treatment has evolved from the standard, relatively nonspecific chemo- and radiotherapy with significant cytotoxic side effects towards immunotherapeutic approaches with selective, target-mechanism-based effects. Until now, several tumor antigens for antibody therapy as well as for cancer vaccination have been identified for different cancer types and they also have shown promising results in clinical trials. Antibody therapies as the most successful form of cancer immunotherapy nowadays already led to some approved treatments for some specific cancer types. Although research in both areas yielded positive results for treatment, the identification of tumor antigens with high immunogenicity remains challenging.

Therefore, we developed an bioinformatics pipeline for the efficient and systematic discovery of overexpressed, combinatorial molecular targets with high prognostic and therapeutic relevance for clinical, immunotherapeutic cancer treatment based on the statistical analysis of RNA-sequencing data. Our approach is able to develop an intelligent, highly-specific description of individual tumor patients within heterogenic cohort collectives leading to immunologically addressable targets in subgroups of patients. We achieve this by actively modeling potential negative side effects of the immunotherapy utilizing expression data of 30 different, normal, cancer-free human tissues. The whole pipeline is based on a database comprising patient and normal tissue data.

First application of our pipeline on breast-cancer data, successfully automatically identifies potential combinatorial markers and targets from RNA-sequencing data, which separate cancer from normal tissue. The markers and targets are currently under subsequent experimental validation.

42. **Carlus Deneke and Bernhard Renard**
Machine learning for virulence prediction

Abstract: In recent years, the number of sequenced and annotated pathogenic organisms has continuously increased. It has become a major goal of bacteriologists to identify the virulence factors, i.e. those sequences that ultimately induce pathogenicity. Since information of large data sets needs to be processed, methods from bioinformatics and statistical learning are in strong demand to support this process.

This contribution reports on the application of machine learning for the classification of virulence-related protein sequences in bacteria. In a first step, we define positive and negative data. For the former, specialty databases are available; however, particular care has to be taken for choosing an appropriate negative set in order to avoid systematic biases. Here, we propose a novel approach where we choose proteins based on gene ontology annotation. From these data, we generate a variety of features based on protein sequences. These features range from simple count statistics of the occurrence of amino acids to more indirectly inferred information related to the protein structure. We aim to define many diverse feature types in order to fully capture the information content of each protein sequence.

We implement various supervised learning strategies - such as random forests and support vector machines - and evaluate their respective advantages. As a result, we obtain sensitivity and specificity well above 80 % on independent test sets. Furthermore, the systematic analysis of the feature importance yields the most relevant features related to virulence which in turn supports the biological interpretation. Additionally, this strategy allows a systematic feature selection - such that less than 50 selected features have equal discriminatory power as the 900 dimensional full feature space.

In conclusion, the approach presented here describes a powerful method to classify previously unknown protein sequences in respect of their potential to affect the virulence of an organism.

43. **Aarif Mohamed Nazeer Batcha and Ulrich Mansmann**
Prevalence of potential mutations in colorectal cancer addressed genes in healthy population

Abstract: Colorectal carcinoma (CRC) or bowel cancer is the third most commonly diagnosed cancer and fourth leading cause of cancer mortality in European Union (EU). During 2012, the incidence and mortality rates of CRC were estimated to be 46.3 per 100,000 person years and 18.4 per 100,000 person years, respectively [1]. The prognosis for patients with CRC is heavily dependent on the stage at diagnosis. Several screening procedures were available to reduce its incidence and mortality rate and some of them were found to be effective. Though a number of risk prediction models have been developed for CRC, a very few genetic models were available [2]. Our interests lie in finding out the prevalence of potential mutations in the CRC addressed genes among healthy population and to develop a genetic prediction model to determine the risk of CRC for the population of EU. The genotype database obtained from kooperative gesundheitsforschung in der region Augsburg (KORA) study will be used for analyses. The task involves development of an algorithm interconnecting the information from KORA database and the variants published by cancer genome atlas [3] for the prediction model using bio-informatics tools such as PLINK. The topic will also include the issues and the practical difficulties faced during the process.

References

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JWW, Comber H, Forman D,Bray F. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer. 2013 Apr;49(6):1374-403. doi: 10.1016/j.ejca.2012.12.027.

2. Win AK, Macinnis RJ, Hopper JL, Jenkins MA. Risk prediction models for colorectal cancer: a review. Cancer Epidemiol Biomarkers Prev 2012;21:398-410.

3. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330-337 (19 July 2012), doi: 10.1038/nature11252.

44. **Khalid Abnaof, Joao Dinis and Holger Fröhlich.**
Using Consensus Clustering to Explore Biological Effect Similarities of Drug Treatments based on Integrated Biological Knowledge from Multiple Sources - An Example Study on HIV and Cancer

Abstract: We asked the question, whether approved drugs for different diseases (here: HIV and cancer) fall into separable clusters according to their induced biological response.

We collected for each known drug target (199 in cancer, 22 in HIV) a multitude of biological information, such as involvement into biological processes, contained protein domain annotation and position in specific pathways. Moreover, drug-target binding affinities (Ki values) were retrieved.

We developed an integrated similarity measure for drugs comparing their expected biological effects. Our biological effect similarity (BES) uses protein target similarities combined from multiple information sources in a probabilistic manner. These similarities are then extended to drug similarities: For each drug pair we weight target similarities with drug binding affinities. The result is one BES measure per drug pair. Using complete linkage consensus clustering we were able to find well separated clusters in HIV and in cancer. Analysis of targets of compound clusters, revealed in most cases a clear enrichment of distinct Gene Ontology terms, KEGG pathways, protein domains and sequence motifs. Our identified clusters could not be found in a traditional ligand-based clustering using fingerprints. This was also true when a joint clustering of HIV and cancer drugs with respect to the BES measure was conducted. In that case we could specifically find clusters containing drugs from both diseases, hence indicating biologically close effects despite of a very different medical application area.

This work suggests the importance of a joint view on the compound-target space. Our proposed BES measure may help to uncover drugs with biologically similar effects and thus help repurpose existing drugs for novel areas.

45. **Rewati Tappu and Daniel Huson**
Using MEGAN5 to analyse medical microbiome data

Abstract: Metagenomics, the study of microbes in the environment using DNA sequencing, depends upon dedicated software tools for processing and analysing very large sequencing datasets. One such tool is MEGAN (MEtaGenome ANalyzer), which can be used for the interactive analysis of one or multiple metagenomic or metatranscriptomic datasets, both taxonomically and functionally.
MEGAN uses the Lowest Common Ancestor (LCA) algorithm for binning of reads that have been compared against a reference database into the different hierarchical levels of NCBI taxonomy while functional analysis is performed by mapping reads to the SEED, COG and KEGG classifications. Multiple samples can be compared by normalisation or sub-sampling from the reads and then can be visualised using a wide range of charting techniques. PCoA analysis, co-occurrence plots and clustering methods allow high level comparison of a large numbers of samples. It also calculates rarefaction curves and diversity indices like the Shannon-Index. Different attributes of the samples can be captured and used within analysis. LCA parameters can be changed and optimised for different datasets. The program supports various input formats for data and can export analysis results in different text based graphical formats. Here we illustrate the use of MEGAN in the context of medical studies involving human microbiome data.

46. **Marc Bonin, Irene Ziska, Jekaterina Kokatjuhha, Pascal Schendel, Karsten Mans, Biljana Smiljanovic, Till Sörensen, Andreas Grützkau, Bruno Stuhlmüller and Thomas Häupl.**
Comparative Analysis between Rheumatoid Arthritis and Arthritis Model: Study of the Functional Components in Expression Profiles of Synovitis

Abstract: Background and objective: Microarray experiments can be used to compare samples of healthy and diseased patients. The objective is to identify up or down regulated genes in the diseased sample which could be the key to understand the disease and develop a treatment. The Problem is that the taken samples often not only consist of one but of multiple different cell types and this composition can greatly differ between the healthy and the diseased sample because of the migration of immune cells. Therefore it is important to distinguish if an expression level of a gene is changed due to regulation or as a result of a change in the cell composition.

Materials and Methods: Transcriptomes of immune cell types including T-cells, B-cells, NK-Cells, monocytes and granulocytes as well as transcriptomes of synovial tissue biopsies from patients with rheumatoid arthritis and osteoarthritis were generated using the Affymetrix platform HG-U133Plus 2.0. The MG-4302 GeneChip transcriptomes of the same panel of cells (GSE6506) and synovial tissues from the collagen induced arthritis model in DBA-1 mice (GSE13071) were selected from the GEO database.

Results: We developed a model to compute cell fractions for each sample by using cell-specific marker genes from reference transcriptomes for each cell type involved. Based on these fractions a virtual profile was computed for each sample that represents the sample-specific mixed cell profile based on the reference transcriptomes and thus without the disease-related regulation of genes. These so called virtual signals were compared to the corresponding real signal to estimated the level of gene regulation. The given model was used to compare the gene regulation in the human disease of rheumatoid arthritis with the collagen induced arthritis model in DBA-1 mice. This new analysis technique revealed activity dependent infiltration of monocytes into the synovial tissue in the DBA-1 mouse model. In contrast to the mouse model, the human disease transcriptomes included patterns of T- cells and B-cells as well as monocytes. With respect to the regulatory changes, there were significant differences not only by quantity but also by quality. For example differences especially in chemokine regulation like CXCL13 between human disease and animal model are in accordance with the observed difference in cell type infiltration.

Conclusion: In summary a model to differentiate between a change in expression level based on regulation and a change due to a different cell composition was developed and successfully applied for synovitis. The data indicate differences between human disease and animal model and can provide important information on the selection of appropriate animal model for the development of therapeutic targets in human diseases.

47. **Marc Bonin, Florian Heyl, Jekaterina Kokatjuhha, Sascha Johannes, Irene Ziska, Pascal Schendel, Karsten Mans, Biljana Smiljanovic, Till Sörensen, Bruno Stuhlmüller and Thomas Häupl**
Identification of Co-Expression Networks of Inflammatory Response in Immune Cells

Abstract: Background and objective: In chronic inflammatory diseases the pathomechanisms of persistence are insufficiently understood. Systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA) are two of the many diseases, which are increasingly common and lead to new diagnostic and therapeutic challenges. IFN-α, TNF-α and pathogen-associated molecular patterns (PAMPs) like lipopolysaccharides (LPS) play key roles in the process of the inflammatory response in immune cells. Especially IFN-α seems to be associated with SLE and implicates the various changes in the persistence of this illness.

Materials and Methods: Transcriptomes induced in monocytes by LPS, TNF-α and IFN-α were used as a reference to identify and screen for networks of known stimuli (GSE38351). Programming was based on R, JAVA and PHP to provide a framework for analysis and storage of data.

Results: In this study, the influence of the three molecular stimuli IFN-α, TNF-α and LPS on CD14-monocytes was investigated. With the help of this data and the co-expression networks as a model, relationships between the genes were analyzed. Thus study was based on an extensive de novo analysis with heuristic filter methods, correlation analyses and hierarchical clustering. A web application was developed to display the correlations in a defined list and to quickly identify relevant genes and functional relationships. With the aid of the application specific networks for two genes (IFI44 and OASL) were identified that cover no more than a hundred genes. Together with an automatically produced intersection between the results of three analyses of three different measurements the networks appear to be highly reproducible. With Pearson correlation, Spearman correlation and mutual information, three different algorithms were applied. Comparison between two normalization methods (the RMA-normalization and a modified version of the MAS5-normalization) confirms the robustness of the networks.

Conclusion: In summary, the study demonstrates that different networks for the individual triggers can be identified. This includes especially networks, which are typical for IFN-α and which are associated with SLE. The data also demonstrate overlapping activities of the different stimuli and that selection of appropriate references for network calculation is critical.

48. **Marc Bonin, Pascal Schendel, Karsten Mans, Florian Heyl, Jekaterina Kokatjuhha, Sascha Johannes, Irene Ziska, Biljana Smiljanovic, Till Sörensen, Bruno Stuhlmüller and Thomas Häupl**

Prediction for Successful Treatment of Methotrexate in Rheumatoid Arthritis with mRNA and miRNA Microarraydata

Abstract: Background and objective: Treatment of chronic arthritis is challenged by the need to adapt dosis or exchange therapeutic agents without prior knowlegde of the individual response characteristics. With genomewide screening of transcriptional activities in whole blood, we hope to identify molcular patterns that help to distinguish responders from non-responders prior to treatment and thus may support therapeutic decisions.

Materials and Methods: Samples were collected before Methotrexate treatment. The transcriptomes were determined by Affymetrix technology and the therapeutic response level by clinical follow-up in an observational study. To select potential molecular predictors and to compare the molecular classifiers between different clinical response groups, various R-packages were applied.

Results: Good to very good predictors could be identified by using the Limma and the Lasso algorithms in the mRNA transcriptoms, which enabled to classify nearly without an error by linear discrimination analysis (LDA). Between groups of genes determined by different selection methods an overlap up to 40% could be reached and hierachical clustering generated nearly perfect grouping. Nevertheless among mRNAs the heatmap patterns seemed to be in part heterogeneous. The analysis was repeated after splitting the samples into two groups with respect to the expression level of the gene HLA-DRB4, a gene locus, which is genetically important for risk prediction in rheumatoid arthritis. Between the molecular predictors of response for the two groups of HLA-DRB4 positiv and negativ patients no overlap could be found. Also the overlap with the predictors of the combined group decreased notable. Similar results were observed when analysing the microRNA transcriptoms. Finally the samples were seperated into a test and a training set for an independent validation. Only the investigation of the groups spitted by HLA criteria showed adequate reproducibility whereas the combined group obviously generated unstable predictors.

Conclusion: In summary, combining the advantages of different algorithms like Limma, Lasso and LDA for selecting and testing molecular predictors for clinical response increases the dignostic power of biomarkers. Nevertheless, appropriate characterization and splitting into distinct subgroups is essential to increase reproducibility and validity in biomarker development.

49. **Birgit Henrich, Madis Rumming, Alexander Sczyrba, Eunike Velleuer, Ralf Dietrich, Michael Gombert, Sebastian Rahn, Wolfgang Gerlach, Jens Stoye, Arndt Borkhardt and Ute Fischer**
Mycoplasma salivarium-colonised oral squamos cell carcinoma

Abstract: We describe the detection of Mycoplasma salivarium, a non-pathogenic commensal, on the surface of a squamous cell carcinoma of the tongue of a patient with Fanconi anaemia (FA). Employing Roche/454 sequencing of 16S rDNA gene amplicons and TaqMan-PCR we analysed the oral microbiome utilizing the QIIME 454-analysis [CJ10] pipeline of this FA patient in comparison to that of a FA patient with a benign leukoplakia and five healthy individuals. The microbiota of the FA patient with leukoplakia correlated well with that of the healthy controls. A dominance of Streptococcus, Veillonella and Neisseria species was typically observed. In contrast, the microbiome of the OSCC bearing FA patient was dominated by Pseudomonas aerug- inosa at the healthy sites, which changed to a predominance of 98% M. salivarium on the tumour surface. Quantification of the mycoplasma load by TaqMan-PCR confirmed the prevalence of M. salivarium at the tumour sites. These new findings suggest that this mycoplasma species with its re- duced coding capacity found ideal breeding grounds at the tumour sites. The oral cavity of all FA patients and especially the tumour site were positive for Candida albicans. It remains to be elucidated whether M. salivarium can be used as a predictive biomarker for tumour development in these patients.

References

[CJ10] et al. Caporaso JG. QIIME allows analysis of high-throughput community sequencing data. In Nat Methods 7:335-336. 2010.

50. **Manabu Sugii, Keisuke Kawano and Hiroshi Matsuno**
Extension of Genetic Toggle Switch with a Mathematical Analysis for Artificial Genetic Circuits

Abstract: Synthetic biology is a study which aims at constructing artificial genetic circuits with the support of computational methods. H. H. McAdams and L. Shapiro were proposed a hybrid modeling approach for analyzing genetic networks, which integrates conventional biochemical kinetic modeling within the framework of a circuit simulation [HL95]. Gardner et al. showed a method to construct a 2-state genetic toggle switch in Escherichia coli [TCJ00]. They set a simple biological phenomenon and estimated the parameters from biological information, formulating a system of differential equations enabling bistable genetic circuit. In the current context of synthetic biology, artificial genetic circuits are designed in the following way: after setting a biological target phenomenon to be investigated, reaction parameter estimations among related molecules are conducted based on the dynamic analyses with mathematical models. Finally a system of biological reactions is developed with these molecules in vivo or in vitro. We propose a new method to design a computational model of a genetic circuit on the framework of logical and dynamical formalisms. We extended a genetic toggle switch from 2-state to 3-state by our proposed procedure. Differential equations for the 3-state genetic toggle switch were produced by extending the ones for the 2-state genetic toggle switch. Furthermore, we examined the mathematical behavior of the dynamic model of the 3-state genetic toggle switch extended from the 2-state genetic toggle switch by our proposed procedure.

References

[HL95] McAdams Harley H. and Shapiro L. Circuit simulation of genetic networks. Science, 269(5224):650-656, 1995.

[TCJ00] Gardner T.S., Cantor C.R., and Collins J.J. Construction of genetic toggle switch in Escherichia coli. Nature, 403:339-342, 2000.

51. **Jennifer Scheidel, Leonie Amstein, Jörg Ackermann, Liliana Schaefer, Ivan Dikic and Ina Koch**
Mathematical model of antibacterial autophagy

Abstract: Antibacterial autophagy is an important mechanism of the innate immunity [SBSD12]. Pathogens, like Salmonella, that invade the host cytosol are targeted with ubiquitin for the autophagic pathway [BSB+ 06]. It is very important to understand this process to find new therapies for invading pathogens, which more and more frequently become resistant to antibiotics. To gain a better understanding of the biological system of antibacterial autophagy, we developed a first mathematical model of this biological process. We applied the Petri net formalism to semi-quantitatively model the processes of Salmonella ubiquitination and the recognition of the autophagy receptors. The current model is based on an extensive literature research and provides a useful basis for the further integration of quantitative data. We used the tool MonaLisa for the construction, the analysis, and the simulation of the Petri net model [EANK13]. The mathematical analysis of the Petri net checks the model for consistency and biologically meaningful behavior.

References

[BSB+06] Cheryl L Birmingham, Adam C Smith, Malina A Bakowski, Tamotsu Yoshimori, and John H Brumell. Autophagy controls Salmonella infection in response to damage to the Salmonella- containing vacuole. Journal of Biological Chemistry, 281(16):11374-11383, 2006.

[EANK13] Jens Einloft, Jörg Ackermann, Joachim Nöthen, and Ina Koch. MonaLisa-visualization and analysis of functional modules in biochemical networks. Bioinformatics, 29(11):1469-1470, 2013.

[SBSD12] Shabnam Shaid, Christian Brandts, Hubert Serve, and Ivan Dikic. Ubiquitination and selective autophagy. Cell Death Differentia- tion, 20(1):21-30, 2012.

52. **Leonie Amstein, Jennifer Scheidel, Jörg Ackermann, Simone Fulda, Ivan Dikic and Ina Koch**
Mathematical model of TNFR1 signal transduction

Abstract: Death receptors such as tumor necrosis factor receptor 1 (TNFR1) control essential cellular processes like cell death, proliferation and inflammation. Thus, it features a strict regulatory network to modulate the cellular response either towards apoptosis, necroptosis or NF-κB activation. Nevertheless, these pathways are often found to be disrupted in cancer cells and other inflammatory diseases [DPHM12]. In order to elucidate these intertwined cellular pathways and their molecular regulation, we apply systems biology approaches. The lack of kinetic data is often a drawback for modeling signal transduction pathways, but Petri nets (PN) provide a suitable tool to model these systems, as they still allow for topological and dynamical network analysis [KRS11]. We develop a PN model that describes the cellular processes of the TNFR1 signaling system, such as formation of macromolecular complexes, gene expression and signaling outcome. Here, we consider the regulation via post-translational modifications like ubiquitiylation [FRD12]. For first verification studies, the PN is analyzed according to its structural properties and their biological significance. If kinetic data is applied, this semi-quantitative model allows for further analysis like stochastic simulations.

References

[DPHM12] Laura Dickens, Ian Powley, Michelle Hughes, and Marion MacFarlane. The complexities of life and death: Death receptor signalling platforms. Experimental Cell Research, 318(11):1269-1277, 2012.

[FRD12] Simone Fulda, Krishna Rajalingam, and Ivan Dikic. Ubiquitylation in immune disorders and cancer: from molecular mechanisms to therapeutic implications. EMBO Molecular Medicine, 4(7):545-556, 2012.

[KRS11] Ina Koch, Wolfgang Reisig, and Falk Schreiber. Modeling in Systems Biology: The Petri Net Approach. Springer Berlin / Heidelberg, 2011.

53. —

54. **Oliver Philipp, Andrea Hamann, Heinz D. Osiewacz and Ina Koch**
The autophagy interaction network of the fungal aging model Podospora anserina

Abstract: Cellular quality control (QC) is a key for proper development and function of any biological system. Impairments in the various QC pathways lead to degeneration, aging, and disease. During the last years, the molecular and cellular pathways involved in the control of aging and lifespan have been studied in the well-established aging model Podospora anserina [OHZ13]. A recent genome-wide longitudinal transcriptome analysis revealed for the first time in P. anserina that autophagy transcripts are up-regulated during aging, while those related to proteasomes and ribosomes decrease in abundance [PHW+13]. In a subsequent study an age-related increase in autophagy was experimentally validated and a longevity assurance function of autophagy was demonstrated [KHP+14]. Unfortunately, the control of autophagy and of the other individual components of QC network and their interactions in respect to aging is poorly understood. Towards this goal, we set out to identify and mathematically model such interactions. In a first step, using published data of yeast and human, we unravel the P. anserina autophagy interactome by means of a new implemented algorithm. This algorithm identifies pairs of P. anserina proteins which are homologous to pairs of known interacting proteins in yeast AND human. Our putative P. anserina autophagy interaction-network includes the most conserved and important components and interactions and consists of more than 100 proteins and interactions comprising initiation, elongation, transport and degradation. Based on this network a mathematical model will be developed to simulate the dynamics of autophagy processes.

References

[KHP+14] Laura Knuppertz, Andrea Hamann, Francesco Pampaloni, Ernst Stelzer, and Heinz D. Osiewacz. Identification of autophagy as a longevity-assurance mechanism in the aging model Podospora anserina. Autophagy, 10:822-834, 2014.

[OHZ13] Heinz D. Osiewacz, Andrea Hamann, and Sandra Zintel. Assessing organismal aging in the filamentous fungus Podospora anserina. Methods Mol Biol, 965:439-462, 2013.

[PHW+13] Oliver Philipp, Andrea Hamann, Alexandra Werner, Ina Koch, and Heinz D. Osiewacz. A genome-wide longitudinal transcriptome analysis of the aging model Podospora anserina. PLoS ONE, 8:e83109, 2013.

55. **Stefan Schuster**
On the history of the Michaelis-Menten equation

Abstract: About 100 years ago, Leonor Michaelis and Maud Menten published (in German) their famous paper ”Die Kinetik der Invertinwirkung” [MM1913], in which they proposed an equation which is nowadays known as Michaelis-Menten kinetics. While this equation is taught to practically all students in biology, biochemistry and medicine, it is not so well-known where Michaelis had worked at that time. Here, an outline of the careers both of Michaelis and Menten are given. Moreover, it is shown that practically the same equation had been derived earlier by French physico-chemist Victor Henri in 1902 [VH02]. He had even been aware of the nowadays often neglected influence of the reaction product on enzyme saturation. We discuss in what respect Michaelis and Menten extended and generalized the approach. The international synergism in the scientific community in the era around 1900 is outlined, which led to the development of the field of enzyme kinetics [DSMC14].

References

[DSMC14] Ute Deichmann, Stefan Schuster, Jean-Pierre Mazat and Athel Cornish-Bowden: Commemorating the 1913 Michaelis-Menten paper ”Die Kinetik der Invertinwirkung”: three perspectives. FEBS Journal, 281:435-463, 2014.

[MM13] Leonor Michaelis and Maud Menten. Die Kinetik der Invertinwirkung, Biochemische Zeitschrift 49:333-369, 1913.

[VH02] Victor Henri. Théorie générale de láction de quelques diastases. Comptes Rendus Hébdomadaires Séances Academie des Sciences 135:916-919, 1902.

56. **Christian Tokarski, Sebastian Vlaic, Jana Schleicher, Reinhard Guthke and Stefan Schuster**
Hepatic Response to Refeeding - From Ketone Bodies to Lipids

Abstract: Despite several important functions of the liver including detoxification of drugs, alcohol, ammonia and other xenobiotics, as well as synthesis of compounds like bile, membrane lipids and proteins, it is in particularly important in energy storage after meals and in providing energy during periods of starvation. In times of low blood glucose level the brain, which cannot metabolize fatty acids, is generating energy additionally to glucose from ketone bodies that are solely produced in the liver. In the liver these are synthesized by degradation of fatty acids via β-oxidation and exported into the blood. Since a high level of ketone bodies in blood for prolonged periods leads to damage of the nervous system, kidney and to a lower blood pH value, it needs to be lowered. Ketone bodies are predominantly utilized in skeletal muscle, heart and brain. The liver itself is only partially able to utilize them since the two important enzymes limiting the ketolysis, succinyl CoAoxoacid transferase (SCOT) and methylacetoacetyl CoA thiolase (MAT), show very low activity in liver tissue. In order to study the role of the liver in ketone body metabolism, we created a literature based minimal model of the human fatty acid metabolism. Published gene expression data from Vollmers et al. was downloaded from GEO (accession number: GSE13093). Since this data captures the molecular changes in response to refeeding mice after a period of starvation, we seeked to observe differential gene expression among the genes included in our minimal model. To investigate the relations in the expression of genes we performed network inference using ExTILAR. Including prior knowledge of diverse types in the inference we obtained a dynamic gene regulatory network that is able to reproduce the observed dynamics. Connecting this network with the minimal model of hepatic fatty acid metabolism, the whole response of the liver to refeeding can be simulated leading to incorporation of ketone bodies into the liver which seem to be metabolized to form cholesterol or fatty acids via high activity of acetoacetyl-CoA synthetase (AACS) and acetyl-CoA acetyltransferase 2 (ACAT2).

57. **Olga Popik, Björn Sommer, Ralf Hofestädt and Vladimir Ivanisenko**
Pathway effciency evaluation based on protein-protein interaction data and protein localizations

Abstract: Intracellular protein localization plays an important role in cell functioning as biological processes are usually distributed over various intracellular compartments. The variance of the reaction rate among different compartments is usually quite high. But how is it possible to computationally estimate the overall reaction rate of a complete metabolic pathway? We suggest that the frequency determination of intracompartmental as well as intercompartmental protein-protein interactions is an appropriate approach to predict the pathway's reaction rate. There are many databases containing data about protein-protein interactions (PPI) and protein's intracellular localizations. ANDSystem [1] integrates protein intracellular localization data and information on PPI, extracted from various databases such as UniProt, IntAct, KEGG, BIND, etc., and scientific publications. By using data from ANDSystem, a PPI frequency matrix was constructed for all pairs of 14 specific human cellular compartments. This matrix can be used to compare the frequency of intracompartmental as well as intercompartmental PPI.

An initial analysis of the matrices revealed that the frequency of intracompartmental PPI occurs more often in comparison with intercompartmental PPI. Then, the human KEGG pathways were evaluated using these matrices, ranked and compared with random pathways.

References

[DIKI12] P. S. Demenkov, T. V. Ivanisenko, N. A. Kolchanov, and V. A. Ivanisenko. ANDVisio: A new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. In silico biology, 11(3):149-161, 2012.

58. **Martin Lewinski, Christoph Brinkrolf and Ralf Hofestädt**
Topological reconstruction and graph analysis of biological interaction networks

Abstract: Biological interaction networks are commonly used to represent processes and relationships in many applications of systems biology. To enable user-specific network reconstruction, an approach that covers data integration, options of network retrieval, network visualization and network filtering is needed. Due to this, we provide a framework which offers topological network reconstruction based on biological interactions together with graph theoretic approaches for the analysis. The interaction data is retrieved from the data warehouse DAWIS M.D. [HKT+10] and stored for performance reasons into a separate graph database. Additional data (e.g. micro array experiments) can be mapped onto the network and is available for analyses, filtering and visualization of the network. By the implementation of common graph algorithms, we provide the possibility to further analyze and filter networks. To identify cohesive patterns [MCRE09], several attributes (graph theoretic and experimental) can be linked together utilizing densely connected bi-clustering. Using this approach we can detect patterns in the network which have specific similarity features and review current experimental data or assist in the decision making process for the next iteration of experiments.

The implementation of the algorithms are available in the systems biology framework VANESA [BJK+14].

References

[BJK+14] Christoph Brinkrolf, Sebastian Jan Janowski, Benjamin Kormeier, Martin Lewinski, Klaus Hippe, Daniela Borck, and Ralf Hofestädt. VANESA - A Software Application for the Visualization and Analysis of Networks in System Biology Applications. Journal of integrative bioinformatics, 11(2):239, 2014.

[HKT+10] Klaus Hippe, Benjamin Kormeier, Thoralf Töpel, Sebastian Janowski, and Ralf Hofestädt. DAWIS-M.D. - A Data Warehouse System for Metabolic Data. In Klaus-Peter Fähnrich and Bogdan Franczyk, editors, GI Jahrestagung (2), volume 176 of LNI, pages 720-725. GI, 2010.

[MCRE09] Flavia Moser, Recep Colak, Arash Rafiey, and Martin Ester. Mining Cohesive Patterns from Graphs with Feature Vectors. In Chid Apte, Haesun Park, Ke Wang, and Mohammad J. Zaki, editors, Proceedings of the 2009 SIAM International Conference on Data Mining, pages 593-604. Society for Industrial and Applied Mathematics, Philadelphia and PA, 2009.

59. **Jana Schleicher, Stefan Schuster and Reinhard Guthke**
Which mechanism determines the zonation of a fatty liver?

Abstract: A serious problem in modern society is the increased incidence of metabolic disorders attributable to dietary habits. In this context, liver diseases characterized by a non-physiological fat accumulation in the cytosol of hepatocytes (fatty liver, steatosis) induced by a high caloric diet reaches a prevalence of above 20 percent in Western countries. These liver diseases are called non-alcoholic-fatty-liver diseases (NAFLD). The mechanisms which determine the increased hepatic fat accumulation are poorly understood partly due to a lack of knowledge regarding the link between structure and function of the liver. Decades ago, first experimental studies revealed prominent heterogeneity in subcellular structure and enzyme activities between hepatocytes. This laid the foundation for the concept of metabolic zonation, which means that different metabolic pathways are active in the hepatocytes of different zones along the blood vessels (sinusoids). Therefore, hepatocytes, depending on their position, show different rates of metabolic pathways of carbohydrate, ammonia, xenobiotic and lipid metabolism. Notably, under a high fat diet the accumulation of triglycerides along the sinusoid is also zonated. Experimental data showed that fat accumulates first in hepatocytes around the central vein (perivenous zone), but the accumulation can extend to hepatocytes around the portal vein and hepatic artery (periportal zone) by an increased supply of dietary fatty acids. To investigate the establishment of a zonated fatty liver, we developed a mathematical model of hepatic fatty acid metabolism (fatty acid uptake, beta-oxidation, triglyceride synthesis and export) in hepatocytes along a blood vessel. The model showed in the first instance a periportal accumulation of triglycerides as a result of the high concentration of fatty acids streaming from the portal vein. The question arose, which mechanism in vivo leads to a perivenous fat storage? We extend our mathematical model to evaluate possible mechanisms leading to the latter case.

60. **Meik Kunz, Muhammad Naseem, Chunguang Liang and Thomas Dandekar**
Probing the Unknowns in Cytokinin-Mediated Immune Defense in Arabidopsis with Systems Biology Approaches

Abstract: Plant hormones involving cytokinin (CK), salicylic acid (SA), jasmonic acid (JA), ethylene (Et), and auxin, gibberellins, and abscisic acid (ABA) are small signaling molecules known to regulate almost every aspect of the plant life cycle including immune responses [1,2]. Moreover, CK regulates not only growth and development but also immunity and has the potential to modulate defense signaling mediated by SA and JA. Recently, enhanced CK level has been shown to increase plant resistance against pathogen infection [2,3,4,5]. However, the underlying mechanisms highlighting its implications in plant immunity are not well understood. To identify hub points of immune interaction mediated by CK signaling upon pathogen infection in Arabidopsis we adapted systems biology approaches. High confidence Arabidopsis Protein-Protein Interaction networks [6] are mapped to changes in CK-mediated gene expression after treatment with Pst DC3000 (GSE6832.I) and Hpa Noco2 [7]. Nodes of the cellular interactome enriched in immune functions were filtered out and their interacting partners reconstituted into sub-networks (method see reference 1). Based on different criteria such as topological parameters, gene expression and specific immunological relevance we identified functional hubs in our immune subnetworks. According to our analysis, de-repression of CK responses through the deletion of type-A ARRs promotes the SA pathway of resistance and points to a link between CK signaling and WRKY transcription factors in controlling immune dynamic in Arabidopsis. In our analysis, enhanced CK levels through external application modulate immune responses by activating JA pathway nodes. Taken together, our analyses identified hubs integrating functional modules as cross-linking agents between CK-mediated immune defense and pathways of resistance against pathogen infection in plants.

References

[1] Naseem, Kunz et al. (2014) Probing the Unknowns in Cytokinin-Mediated Immune Defense in Arabidopsis with Systems Biology Approaches. Bioinformatics and Biology Insights 2014:8 35-44. doi:10.4137/BBI.S13462.

[2] Kunz, M et al, (2013) Hormone Signaling Networks Open Multiple Routes for Immunity and Disease in Plants. Biohelikon: Immunity and Diseases 2013 1:2.

[3] Choi, J. et al. (2011) Cytokinins and plant immunity: old foes or new friends? Trends Plant Sci. 16, 388-394.

[4] Naseem, M. et al. (2012) Integrated systems view on networking by hormones in Arabidopsis immunity reveals multiple crosstalk for cytokinin. The Plant cell 24, 1793-814.

[5] Navarro, L et al. (2008) DELLAs control plant immune responses by modulating the balance of jasmonic acid and salicylic acid signaling. Curr Biol 2008; 18:650-5; PMID:18450451.

[6] Szklarczyk, D et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011;39 (Database issue):D561-8.

[7] Argueso, C.T. et al. (2012) Two-component elements mediate interactions between cytokinin and salicylic acid in plant immunity. PLoS genetics 8:e1002448.

61. **Thomas Wiebringhaus and Heinrich Brinck**
Intermodular Connections Based On Highly Connected Regions In The Human Proteome

Abstract: Biological networks have evolved to fulfill different cellular tasks and so their elucidation will give more insight into the complex organization of the eukaryotic cell [Wiebringhaus T. (2013)]. Recently, the bridging centrality parameter was developed for finding intermodular connections by establishing a combined measurement of the betweenness centrality (BC) [Goh et al., 2003] and the bridging coefficient, which measures how well a node is located between highly connected nodes, and thus global and local parameters are combined. While the BC finds bottlenecks based on information flow, the bridging coefficient adds local topological parameters by means of node connectivity [Hwang et al., 2006]. In this study solely regional parameters should be considered for finding intermodular connections. As a new approach, a method called Interconnectedness Coefficient (IC) based on proteins located between highly clustered and not highly connected regions is introduced. Shortly, the IC describes the clustering of a node related to the average clustering of the direct neighbours as a node weight. Interestingly, detailed manual curation suggests that the parameter detects proteins between highly interwired complexes and underlines that some modules are linked via other mediating modules.

References

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

[Hwang W. et al. (2006)] Bridging Centrality: Identifying Bridging nodes in scale-free networks. KDD 2006 August 2023, Philadelphia. PA, USA.

[Wiebringhaus T. (2013)] High Betweenness- Low Connectivity (HBLC) Signatures in the Human Proteome GCB 2013 P79

62. **Jan Grau and Jens Keilwagen**
Discriminative modeling of dependencies in DNA binding sites

Abstract: Binding of transcription factors to the DNA is one of the keystones of transcriptional gene regulation. While the existence of of statistical dependencies between binding site positions is widely accepted, their relevance for computational predictions has been debated. For existing dependency models like Bayesian trees, we need to distinguish between the discrete space of possible dependency structures and the continuous space of model parameters, which makes such models unusable for numerical optimization required, e.g., by many discriminative learning principles. To overcome this issue, we propose sparse local inhomogeneous mixture (Slim) model that allow for learning dependency structure and model parameters simultaneously. Instead of discrete structures, Slim models employ soft feature selection, which combines putative dependency structures in a weighted manner. We find that Slim models yield a substantially better prediction performance than previous models on the genomic context PBM data sets of Mordelet et al. [MHH+13]. We also apply Slim models to a large collection of ChIP-seq data sets from the ENCODE project and compare their performance to standard position weight matrices and weight array matrix models in the common motif discovery framework of Dimont [GPGK13]. We observe that for the large majority of data sets, Slim models yield an improved prediction performance compared to the other models. The dependency structures responsible for the improved performance are highly diverse ranging from broad heterogeneities to sparse dependencies between neighboring and non-neighboring binding site positions.

References

[GPGK13] Jan Grau, Stefan Posch, Ivo Grosse, and Jens Keilwagen. A general approach for discriminative de novo motif discovery from highthroughput data. Nucleic Acids Research, 41(21):e197, 2013.

[MHH+13] Fantine Mordelet, John Horton, Alexander J. Hartemink, Barbara E. Engelhardt, and Raluca Gordan. Stability selection for regression-based models of transcription factor-DNA binding specificity. Bioinformatics, 29(13):i117-i125, 2013.

63. **Thomas Sütterlin, Kai Safferling and Niels Grabe**
EPISIM: A user-friendly platform for multiscale multicellular modeling and simulation of biological systems

Abstract: Multi-scale in silico models are increasingly developed to obtain comprehensive mechanistic insight into biological systems. The creation and by that the implementation of such models is increasingly complex and gives rise to a demand for dedicated user-friendly modeling and simulation tools. To this end, we developed the EPISIM platform for graphical multi-scale modeling and simulation of multicellular systems. [SKD+13] Each EPISIM tissue model is subdivided in a cell behavioral and a biomechanical model (CBM and BM). The BM covers all spatial and biophysical cell properties. Different BMs (lattice and off-lattice) are offered by EPISIM. A BM is dynamically linked to a CBM. Discrete (deterministic and/or stochastic) and continuous models on cellular on subcellular scale can be combined in a CBM by semantically integrating quantitative subcellular models based on the Systems Biology Markup Language (SBML) standard. Finally, reaction-diffusion models of e.g. cytokines or chemokines can be integrated in a multi-scale tissue model with extracellular diffusion fields. CBMs are graphically modeled with process diagrams and automatically translated in to executable code in the EPISIM modeling system. The EPISIM simulation environment conducts an agent-based tissue simulation based on this code. We realized a multi-scaled model of human epidermal homeostasis by integrating Tyson's cell cycle model into a CBM of keratinocyte proliferation and differentiation. Simulation yields a characteristically stratified epidermis with typical tissue kinetics. With a second keratinocyte model we were able to reproduce a novel wound repair mechanism in silico. This mechanism was previously revealed by experiments with our standardized in vitro wound model based on full thickness epidermal tissue cultures. [SSW+13]

References

[SKD+13] Thomas Stterlin, Christoph Kolb, Hartmut Dickhaus, Dirk Jger, and Niels Grabe. Bridging the scales: semantic integration of quantitative SBML in graphical multi-cellular models and simulations with EPISIM and COPASI. Bioinformatics (Oxford, England), 29(2):223-9, January 2013.

[SSW+13] Kai Safferling, Thomas Stterlin, Kathi Westphal, Claudia Ernst, Kai Breuhahn, Merlin James, Dirk Jger, Niels Halama, and Niels Grabe. Wound healing revised: A novel reepithelialization mechanism revealed by in vitro and in silico models. The Journal of cell biology, 203(4):691-709, November 2013.

64. **Rim Zaripov**
Approaches to harnessing the complexity of metaproteomic data

Abstract: Complex microbial communities are an integral part of the human body. Determination of microbiome's diversity is important for understanding its role in maintaining the host organism's health and for finding ways to regulate microbial community preventing and treating diseases. Modern analysis of microbiome includes multiple complementary meta'omic approaches that enable fully and accurately characterize the microbial communities and their interactions with the environment and the host organism. In this study, we try to choose the best reference metaproteome database to explore the diversity of microbiome. After obtaining spectra of peptides from a mass-spectrometer we searched these peptides on primary protein sequences of all variants of selected databases. A number of challenges, such as the exact identification of each individual peptide, appears because of shortcoming of metaproteome reference database along with the conservatism of some proteins [1]. To assess the quality of protein identification offered some parameters (P1-P7). To improve the quality of the reference database and identification of peptides this study suggests the following approaches: 1) the use of different reference meta proteomic databases: bgi, SwissProt and ”triple” database (bgi + SwissProt + E-coli); 2) narrowing reference database according to metagenome; 3) correcting the reference database with single amino acid polymorphisms (SAAPs). Moreover a comparison of metaproteome-based phylogenetic tree vs metagenome-based phylogenetic tree was made to find out if they are alike, for a case we have only one of these sets. It is important not only to study completeness of chosen DBs, but to study their complexity, which depends on a number of well-identified bacteria, a number of bacteria that peptides was sorted with and etc. To asses this complexity we created a set of other parameters (P'1 - P'4).

References

[1] Muth, Thilo, et al. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. Molecular BioSystems,9.4: 578-585, 2013

65. **Heiko Giese, Joerg Ackermann, Ilka Wittig, Ulrich Brandt and Ina Koch**
Exemplary evaluation of complexome profiling data using NOVA

Abstract: Motivation: The isolation of large native macromolecular complexes, identification of components and their dynamics, is a difficult task, requiring advanced proteomic strategies like complexome profiling [HBS+12]. Complexome profiling uses blue-native electrophoresis (BNE) to separate protein mixtures.

Methods: Proteins that are subunits of the same complex are expected to have similar migration profiles which are measured by label-free quantitative mass spectrometry. We developed NOVA - a flexible, interactive tool for the analysis and visual inspection of complexome profiling data [GAH+14].

Results: We analyze a complexome profiling of rat heart mitochondria [HBS+12]. Focusing on protein migration profiles of known subunits of the oxidative phosphorylation (OXPHOS) complexes we will illustrate the basic steps for the evaluation of complexome profiling data. We show that NDUFA4, a previously known subunit of the respiratory chain complex I, is actually a subunit of complex IV as described by Balsa et al., 2012 [BMPC+12].

References

[BMPC+12] Eduardo Balsa, Ricardo Marco, Ester Perales-Clemente, Radek Szklarczyk, Enrique Calvo, Manuel O. Landázuri, and José A. Enríquez. NDUFA4 Is a Subunit of Complex IV of the Mammalian Electron Transport Chain. Cell Metabolism, 16(3):378-386, August 2012.

[GAH+14] Heiko Giese, Jörg Ackermann, Heinrich Heide, Lea Bleier, Stefan Dröse, Ilka Wittig, Ulrich Brandt, and Ina Koch. NOVA: a software to analyze complexome profiling data. submitted, 2014.

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

66. **Ralf Eggeling, Andre Gohr, Jens Keilwagen, Michaela Mohr, Stefan Posch, Andrew D. Smith and Ivo Grosse**
On the Value of Intra-Motif Dependencies of Human Insulator Protein CTCF

Abstract: The binding affinity of DNA-binding proteins such as transcription factors is mainly determined by the base composition of the corresponding binding site on the DNA strand. Most proteins do not bind only a single sequence, but rather a set of sequences, which may be modeled by a sequence motif. Algorithms for de novo motif discovery differ in their promoter models, learning approaches, and other aspects, but typically use the statistically simple position weight matrix model for the motif, which assumes statistical independence among all nucleotides. However, there is no clear justification for that assumption, leading to an ongoing debate about the importance of modeling dependencies between nucleotides within binding sites. In the past, modeling statistical dependencies within binding sites has been hampered by the problem of limited data. With the rise of high-throughput technologies such as ChIP-seq, this situation has now changed, making it possible to make use of statistical dependencies effectively. In this work, we investigate the presence of statistical dependencies in binding sites of the human enhancer-blocking insulator protein CTCF by using the recently developed model class of inhomogeneous parsimonious Markov models, which is capable of modeling complex dependencies while avoiding overfitting. These findings lead to a more detailed characterization of the CTCF binding motif, which is only poorly represented by independent nucleotide frequencies at several positions, predominantly at the 3' end.

67. Axel Rasche, Matthias Lienhard and Ralf Herwig

ARH/ARH-seq: Discovery Tool for Differential Splicing in High-throughput Data

Abstract: Alternative splicing (AS) is a key mechanism for generating the complex proteome of an organism. AS has been observed within a variety of biological conditions, for example, in tissue expression, with respect to human diseases and in protein modification. The computational prediction of alternative splicing from high-throughput data is inherently difficult and necessitates robust statistical measures because the differential splicing signal is overlaid by influencing factors such as gene expression differences and simultaneous expression of multiple isoforms amongst others. We propose ARH, a discovery tool for differential splicing in case control studies that is based on the information-theoretic concept of entropy. ARH-seq works on high-throughput sequencing data [RLY+14] and is an extension of the ARH method that was originally developed for exon microarrays [RH09]. We show that the method has inherent features, such as independence of transcript exon number and independence of differential expression, what makes it particularly suited for detecting alternative splicing events from sequencing data. In order to test and validate our workflow we challenged it with publicly available sequencing data derived from human tissues and conducted a comparison with eight alternative computational methods. In order to judge the performance of the different methods we constructed a benchmark data set of true positive splicing events across different tissues agglomerated from public databases and show that ARH is an accurate, computationally fast and high-performing method for detecting differential splicing events.

References

[RH09] Axel Rasche and Ralf Herwig. ARH: Predicting Splice Variants from Genome-wide Data with Modified Entropy. Bioinformatics, 26:84 90, 2009. (OpenAccess).

[RLY+14] Axel Rasche, Matthias Lienhard, Marie-Laure Yaspo, Hans Lehrach, and Ralf Herwig. ARH-seq: identification of differential splicing in RNA-seq data. Nucleic Acids Research, 2014. (accepted).

68. **Cedric Saule and Robert Giegerich**
Observations on the Feasibility of Exact Pareto Optimization with Applications to RNA folding

Abstract: Pareto optimization combines independent objectives by computing the Pareto front of its search space, defined as the set of all candidates for which no other candidate scores better under both objectives. This gives, in a precise sense, better information than an artificial amalgamation of different scores into a single objective, but is more costly to compute. We define a general Pareto product operator $*_{Par}$ on scoring schemes. Independent of a particular algorithm, we prove that for two scoring schemes A and B used in dynamic programming, the scoring scheme $A *_{Par} B$ correctly performs Pareto optimization over the same search space. We show that a "Pareto-eager" implementation of dynamic programming can achieve the same asymptotics as a single-objective optimization which computes the same number of results. For RNA structure prediction under the minimum free energy versus the maximum expected accuracy model, we show that the empirical size of the Pareto front remains within reasonable bounds. Without artificial amalgamation of objectives, and with no heuristics involved, Pareto optimization is faster than computing the same number of answers separately for each objective.

69. **Cristina Della Beffa and Frank Klawonn**
Certainty intervals for fold-changes of RNA expression

Abstract: A fold change in RNA expression represents an increase of expression between two conditions. This fold change is generally computed as the ratio of the means between the replicated samples of two conditions. A ratio equal to one means that the corresponding gene is equally expressed in the two conditions, while a very elevated value characterizes highly regulated genes, associated to the respective RNAs. One would expect to get a small number of very high ratios, due to the biological fact that most of the transcripts are generally not regulated. Only the strongly expressed genes are interesting and will be more extensively analysed. Instead, usually those ratios mostly show a wide range, from moderately to highly elevated values, due to the presence of a large number of missing/zero values and many low, not meaningful values. We are developing a statistical approach to correct the fold changes related to RNA sequencing data, computing an interval where the probability of finding the true fold change is very high. This interval is also meant to set a threshold over which the fold ratio is considered ”large enough", to be able to easily exclude a considerable number of not significant transcripts from further analyses. This new Bayesian method aims at shrinking the aforementioned fold changes, especially to decrease those related to genes that are actually poorly expressed in both conditions, to reduce the number of regulated genes, highlighting the few strongly significant expressed ones.

70. **Jens Einloft, Joerg Ackermann and Ina Koch**
Topological properties of metabolic networks

Abstract: Topological properties of metabolic networks can tell us more about the global and local structure and complexity of this networks. To analyze such properties, we analyzed a set of 2641 whole genome metabolism models published in the path2models database [BRS+13]. In order to define the properties for both, reactions and metabolites, we represent those models as directed bipartite graphs. Here, we choose the Petri net formalism [KRS11] to represent the metabolic models. The number of metabolites is strongly correlated with number of reactions, typically a metabolic model has 29% more metabolites than reactions. We computed the distributions of node degree, cluster coefficient, and shortest path length separately for reactions and metabolites in each model. We discuss typical statistical properties of metabolic networks by presenting consensus distributions of the models. For instance, the consensus distribution d(k) of the node degree k determines the probability to find a reaction with k different participating metabolites. The reaction is selected randomly from the set of all reactions in all models. The participating metabolites can be substrates or products of the reaction. Additional valuable information arise from the distinction between in and out degree. Node degree and cluster coefficient of metabolites demonstrated scale freeness and small world property of the networks. We applied box plot statistics to compute outliers in the node degree distribution of metabolites of each model. We assigned the outliers to metabolic hubs. The deletion of these hubs (i.e., secondary metabolites as ATP) has dramatic influence on the network properties. Surprisingly, the reduced networks remain to show the small world property. A completely different picture appears when we turn to reactions. For reactions, the distribution of node degree was much narrower but exhibited a distinct structure. The metabolic models favored reactions with node degree 2 and 5, but reactions with node degree 3 or > 8 are rare.

71. **Norma J. Wendel, Michael U. Höfer, Friedrich Felsenstein, Maria Rosenhauer, Jan Petersen and Antje Krause**
A systems biology approach to herbicide resistance in blackgrass

Abstract: Blackgrass (Alopecurus myosuroides) is one of the most widespread damaging weed in wheat crops in Western Europe. Most populations show resistance to a wide range of herbicides. Although several target site resistance (TSR) mechanisms are already well understood the molecular mechanisms of non-target site resistance (NTSR) are still challenging. In order to determine candidate genes for metabolic herbicide resistance in blackgrass we performed a transcriptome sequencing prepared from control and herbicide treated plants at different time points (4h, 8h and 25h after treatment) from a metabolic resistant biotype [HFRP14]. Due to the limited number of completely sequenced plant genomes, i.e. grass species, we started by mapping the reads to currently available genomes from the ENSEMBL plants genome database (http://plants.ensembl.org/), e.g. Arabidopsis thaliana, Oryza sativa, Aegilops tauschii, Sorghum bicolor and Brachypodium distachyon. The mapping was performed using novoalign (http://www.novocraft.com/) and followed by annotation with HTSeq [APH14]. A systems biology approach was choosen to identify involved metabolic pathways and understand the metabolic background of herbicide resistance.

References

[APH14] S. Anders, P.T. Pyl, and W. Huber. HTSeq - A Python framework to work with high-throughput sequencing data. bioRxiv, 2014.

[HFRP14] M. Höfer, F. Felsenstein, M. Rosenhauer, and J. Petersen. Molekulare Analyse der metabolischen Resistenz in Acker-Fuchsschwanz. Julius-Kühn-Archiv, (443), 2014.

72. **Franziska Metge and Christoph Dieterich**
New Insights to Protein Occupancy Profiling on mRNA

Abstract: RNA binding proteins (RBPs) play an essential role in post transcriptional gene regulation. A disruption of this process is associated with a number of diseases. To understand disease causing aberrations it is necessary to identify protein binding sites. One powerful method to identify binding sites of a single RBP is by photoactivatable ribonucleoside-enhanced UV crosslinking and immunoprecipitation (PAR-CLIP). TC/AG conversions indicate the interaction of a protein and RNA. In parallel to this, popomR has been established to capture the protein occupancy of the transcriptome by crosslinking and oligo(dT) enrichment of 4-thiouridine labelled mRNA [Bea12]. The POPPI pipeline [Sea14] was developed to analyse occupancy profiles based on sequencing data derived from popomR. POPPI detects differentially occupied positions in the transcriptome under different conditions using replicates. Our current work aims to improve POPPI and to learn more about protein occupancy profiles of polyadenylated RNA. We adapted the pipeline to enhance speed and as well as mappability. The signal of RNA-expression is now integrated into the calculation of differentially occupied positions instead of just acting as a cut-off. Additionally, we studied the individual profiles of HEK293 and MCF7 cell-lines to gain greater insight to the nature of popomR profiles and the signal of individual RBPs. With the changes to the pipeline we were able to enhance detectability of differentially occupied binding sites. The deeper insights to occupancy profiles enabled us to identify individual binding sites of single conditions with greater reliability.

References

[Bea12] AG Baltz et al. Molecular cell, June 2012.

[Sea14] M Schueler et al. Genome Biology, January 2014.

73. **Heiner Klingenberg and Peter Meinicke**
Phylogenetic and functional classification of metatranscriptomic sequencing reads

Abstract: Metagenomic sequencing can be used to characterize the entire metabolic potential of a microbial community. In metatranscriptomics, RNA-seq data provides a snapshot of the actual activities of the community in reaction to environmental conditions which can for instance elucidate external influences on microbial life with regard to available C-sources. The data produced by next generation sequencing usually comprise large amounts of short reads. The challenge for computational pipelines in now to efficiently assign taxonomic and functional categories to these reads. With UProC (http://uproc.gobics.de/) a fast protein sequence classification tool is available, which for short reads (100 bp) shows a higher sensitivity than profile-based methods. The UProC tool is based on a dictionary of functionally labeled protein words that are used for classification of a query sequence. Including taxonomic information into the UProC database makes it possible to perform phylogenetic and functional classification at the same time. In addition to the functional category, each word in the dictionary, if possible, is also assigned a specific taxonomic rank, which can range from species to superkingdom. For classification of the query sequence the taxonomic labels of all matching words are combined with an algorithm similar to the lowest common ancestor scheme as applied to BLAST hits. In addition to the Pfam domain-based database, we also provide a KEGG-based UProC database that directly supports the analysis of a metatranscriptome in terms of active metabolic pathways.

74. **Mohammad Tagi Hasanzada and Laila Hassanzada**
An Evolutionary relationship between MicroRNAs involved in MiRNA based Cellular Reprogramming

Abstract: Recently, the Morrisey and Mori groups reported that human and mouse somatic cells can be reprogrammed to produce induced pluripotent stem cells by expressing microRNAs,completely eliminating the need for ectopic protein expression, and other study by Robert Blelloch that focused in promoting the dedifferentiation of somatic cells to iPS cells showed that MicroRNAs have great roles in cellular reprogramming compared with Yamanakas Transcription Factors (OCT4, SOX2, C-MYC, and KLF4) in Cell reprogramming. Therefore, we used bioinformatics techniques, sequence analyses and phylogenetic tree algorithms to understand the evolutionary relationships of human and Mouse MicroRNAs involved in MiRNA based iPS cell generation. We selected fourteen Pri-miRNA genes(miR302/367,miR290-295 and miR371-373 cluster) that are reported by Morrisey, Mori, Blelloch and Houbaviy groups. The Primary MicroRNA gene sequences are collected from miR-base and NCBI databeses. These scores results that there are closely conservations between Primary MiRNA genes between human and mouse species.

75. **Pavankumar Videm, Dominic Rose, Fabrizio Costa and Rolf Backofen**
BlockClust: efficient clustering and classification of non-coding RNAs from short read RNA-seq profiles.

Abstract: Non-coding RNAs play a vital role in many cellular processes such as RNA splicing, translation, gene regulation. However the vast majority of ncRNAs still have no functional annotation. One prominent approach for putative function assignment is clustering of transcripts according to sequence and secondary structure. However sequence information is changed by post-transcriptional modifications [FLSH11], and secondary structure is only a proxy for the true three dimensional conformation of the RNA polymer. A different type of information that does not suffer from these issues and that can be used for the detection of RNA classes, is the pattern of processing and its traces in small RNA-seq reads data.

Here we introduce BlockClust [VRCB14], an efficient approach to detect transcripts with similar processing patterns. We propose a novel way to encode expression profiles in compact discrete structures, which can then be processed using fast graph kernel techniques. We perform both unsupervised clustering and develop family specific discriminative models; finally we show how the proposed approach is both scalable, accurate and robust across different organisms, tissues and cell lines.

BlockClust was successfully applied on a comprehensive set of eukaryotic data. It is the first tool of its kind, which is easily installable and usable on galaxy framework. A complete workflow of BlockClust and its tool dependencies are available at Galaxy ToolShed.

References

[FLSH11] Sven Findeiss, David Langenberger, Peter F. Stadler, and Steve Hoffmann. Traces of post-transcriptional RNA modifications in deep sequencing data. Biol Chem, 392(4):305-13, 2011.

[VRCB14] Pavankumar Videm, Dominic Rose, Fabrizio Costa, and Rolf Backofen. BlockClust: efficient clustering and classification of non- coding RNAs from short read RNA-seq profiles. Bioinformatics, 30(12):i274-i282, 2014.

76. **Preethy Sasidharan Nair, Anju Philips, Harri Lähdesmäki and Irma Järvelä**
RNA-Seq to study music perception

Abstract: Music can evoke emotions and brain imaging studies show that active music listening has effects on brain structure and functions including enlargement of some areas of the brain. But the molecular mechanisms and biological pathways underlying music perception remain unknown. This study aims to fill this gap by investigating changes at the gene level upon passive music listening (traditional European concert) and by to investigate its biological significance in human physiology. Perepheral blood samples were collected before and after concert using PAX gene tubes from seven enthusiastic participants of the Academy of Finland funded MUSGEN project. Total RNA was extracted using Ambion's MagMax kit and checked for integrity using BioAnalyser. Sequencing libraries were prepared from the total RNA using NEBNext® Ultra Directional RNA Library Prep Kit for Illumina and sequenced using HiSeq2500. The sequencing reads are quality controlled by FastQC and aligned to human genome (hg19) using the spliced aligner TopHat2. Annotations from Ensembl (release 75) will be used to assign the mapped reads to genes and differential expression analyses of the RNA-Seq data will be performed using edgeR (R/Bioconductor). Biological pathways involving the differentially expressed genes which are of relevant in human physiology will be investigated with Ingenuity Pathway Analyses. This study is expected to come up with genes and the functional pathways mediating the effect of passive music listening. In addition, comparison of the results with the gene expression changes for a lesser duration of music listening (20 minutes) can help to identify the effect of duration of passive music listening and its beneficiary role in human cognition, learning memory and auditory pathways.

77. **Andrea Tanzer, Ronny Lorenz and Ivo Hofacker**
Detection of RNA G-Quadruplexes within and across Genomes

Abstract: Guanosin-rich RNAs matching the motif $G_L N_{l1} G_L N_{l2} G_L N_{l3} G_L$, with $2 \leq L \leq 7, 1 \leq lx \leq 7$ and $N = A, C, G, U$ fold into locally stable structural elements known as G-quadruplexes. Various scoring algorithms exist that aim to find stable putative quadruplex forming sequences (PQS) based on sequence context only. However, these algorithms result in large amounts of false positives and/or false negatives. Here, we present a new, more reliable approach to score PQS based on thermodynamic local, global, and consensus secondary structure predictions [LBQ+13]. Therefore, our method incorporates the competition between canonical secondary structures and the formation of quadruplexes. For benchmarking, we gathered about 30 experimentally tested human PQS containing 5'UTRs. To investigate evolutionary conservation of PQS within vertebrates we analyzed genome wide alignments (source: genomes.ucsc.edu). Additional alignments were constructed from our benchmarking set with homologs from other animals. Established tools for detection of conserved structures, e.g. RNAz, cannot be applied here due to the lack of G-quadruplex containing training data. Therefore, we screened the human genome for locally stable G-quadruplexes with RNALfold, and analyzed corresponding genome-wide subalignments for conservation with RNAalifold. We refined alignments by discarding sequences of low pairwise similarity, and applied a windowing approach for those exceeding 200 nt. A substantial fraction of locally stable G-quadruplexes in human are con- served across various vertebrate genomes. These findings are confirmed by the alignments derived from out benchmarking set.

References

[LBQ+13] R. Lorenz, S. Bernhart, J. Qin, C. Höner zu Siederdissen, A. Tanzer, F. Amman, I. Hofacker, and P. Stadler. 2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, PP(99):1, 2013.

78. **Abul Islam, Nuria Lopez-Bigas and Elizaveta Benevolenskaya**
Variations in KDM5A/JARID1A/RBP2 Isoform Specific Locations Reveals Contribution of Chromatin-Interacting PHD Domain in Protein Recruitment to Binding Sites

Abstract: RBP2 has shown to be important for cell differentiation control through epigenetic mechanism. The main aim of the present study is genome-wide location analysis of human RBP2 isoforms that differ in a histone-binding domain by ChIPseq. It is conceivable that the larger isoform (LI) of RBP2, which contains a specific H3K4me3 interacting domain, differs from the smaller isoform (SI) in genomic location, may account for the observed diversity in RBP2 function. To distinguish the two RBP2 isoforms, we used the fact that the SI lacks the C-terminal PHD domain and hence used the antibodies detecting both RBP2 isoforms (AI) through a common central domain, and the antibodies detecting only LI but not SI, through a C-terminal PHD domain. Overall our analysis suggests that RBP2 occupies about 77 nucleotides and binds GC rich motifs of active genes, does not bind to centromere, telomere or enhancer regions, and binding sites are conserved compare to random. A striking difference between the only-SI and only-LI is that a large number of only-SI peaks are located in CpG islands and close to TSS compared to only-LI peaks. Enrichment analysis of the related genes indicates that several oncogenic pathways and metabolic pathways/processes are significantly enriched among only-SI/AI targets, but not LI/only-LI peak's targets.

79. **Vaibhav Sabale and Arun Ingale**
Homology modeling and docking study of 3 oxoacyl (acyl carrier protein) synthase II protein of Neisseria meningitidis.

Abstract: Meningitis is inflammation of the protective membrane covering the brain and spinal cord known collectively as the meninges. The bacterial meningitis disease has repeatedly cause outbreak worldwide. Neisseria meningitidis (NM) is major causative agent of bacterial meningitis. The "3 oxoacyl (acyl carrier protein) synthase II" (Beta-Ketoacyl Acp synthase II) enzyme which involved in fatty acid biosynthesis of Neisseria meningitidis. This enzyme in fatty acid synthesis and target for discovery of novel antibacterial agent. In this study Insilico analysis was done by using bioinformatics tools and software. Modeller (V 9.12) use for the 3D structure modeling of 3 oxoacyl (acyl carrier protein) synthase II using wild type E. coli FabF (PDB ID-2GFW) protein 3D structure. The quality and validation of model obtained was performed using the structural Analysis and Verification Server (SAVES). Drugbank database used for the ligand selection. Hex online server and Argus Lab software use for the DOCKING of Protein 3D structure and ligand compound. Study for interaction between protein and ligand Molegro and Gromacs is used. The docking study data showed good fit Root Mean Square Difference (RMSD). This study aims to understand functional aspect to development of novel drug against Meningitis using Insilico approach.

80. **Markus Weisser**
Exchangeable HELM as a new standard for biomolecule representation

Abstract: The Hierarchical Editing Language for Macromolecules (HELM) is an emerging notation standard for the representation of a wide range of biomolecules (e.g. proteins, nucleotides, antibody drug conjugates). The HELM standard was originally developed by Pfizer [ZLX+12] in 2012. The Pistoia Alliance formalized the HELM notation as an open standard and released software tools to work with HELM to the Open Source community.

The hierarchical notation of HELM represents complex macromolecules as polymeric structures of monomers. Monomers can be amino acids and nucleotides as well as unnatural components and chemical modifications. Previously, monomers were represented as their IUPAC amino acid codes or nucleotide codes. Unnatural monomers require a definition in separate databases within the organisations.

We recently developed an extension to the HELM notation to overcome this limitation. The Exchangeable HELM notation (XHELM) is a self-containing format including the structure definition of its component monomers into a single, portable file. This extension allows close collaboration across individual organisations and is a major step for utilizing all the potential of the HELM notation in biomolecular research.

quattro research developed this extension in close collaboration with the Pistoia Alliance. Recently, the Pistoia Alliance released XHELM as an official extension to the HELM notation. The available open source HELM tools fully support this new data format.

References

[ZLX+12] Tianhong Zhang, Hongli Li, Hualin Xi, Robert V. Stanton, and Sergio H. Rotstein. HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. Journal of Chemical Information and Modeling, 52(10):2796-2806, 2012.

81. —

82. **Sabrina Ellenberger, Stefan Schuster and Johannes Wöstemeyer**
Structure modelling, protein-ligand, and protein-protein interaction of sex pheromone processing dehydrogenases in Mucor-like fungi

Abstract: Zygomycetous fungi use trisporic acid and some of its precursors as mediators of sexual development. 4-Dihydromethyltrisporate dehydrogenase (TSP1) is one of the enzymes of the trisporic acid biosynthesis pathway in these fungi. To find out more about this complex communication system, which does not only play a role in sex, but also in parasitic interactions, we created models of tertiary structures and made docking studies with the resulting protein structures. We found one TSP1 sequence for Lichtheimia hyalospora, Phycomyces blakesleeanus, Rhizopus microsporus, Rhizopus oryzae, and Umbelopsis ramanniana.

Against our expectation, we found two different TSP1 sequences in Backusella circina, Mucor circinelloides, and Rhizomucor pusillus. The difference between the two variants can be seen at the C-terminus of the sequences by the occurrence of a short helix in just one of the structures. In one type of the TSP1 proteins, the trisporoid enters the protein diagonally from the right side of the active site like in R. oryzae, while in the second TSP1-type the conformation of the trisporoid ligand is comparable to P. blakesleeanus. Here, the trisporoid enters the protein diagonally from the left side of the active site. Also the preference for a certain series of trisporoids is different between the two TSP1 sequences. The Mucor-like protein prefers trisporoids of the B- and the second, Phycomyces-like TSP1 prefers trisporoids of E-series.

Structure modelling and COTH server predictions support the hypothesis that binding of additional subunits or the formation of dimers are responsible for species specific communication. In the dimer interface region we can find the deciding difference between TSP1 sequences and the template structure, xylose reductase from Candida tenuis. TSP1 sequences contain a glutamine instead of the Ala-173 and a threonine instead of the Arg-180.

83. **Clemens Thoelken**
Xlinq: Identification and Quantification of Chemical Cross-Linked Peptides in Mass Spectrometry Data

Abstract: Chemical cross-linking of proteins in combination with high-resolution mass spectrometry (MS) promises new insights into structural proteomics. While there are several software suites for the identification of linear peptides, the concept of non-linear cross-linked peptides is not addressed by these approaches. The Xlinq software package aims to provide methods for the identification of protein-protein-interactions supporting a variety of cross-linkers with a wide range of properties, leveraging the advantages of complementary detection and fragmentation techniques in MS analysis and validating the results on behalf of structural plausibility and quantitative information. Xlinq follows the XDB approach described by [MCB+07] and provides a fully integrated solution for the identification of potential cross-linked peptide pairs, including a cross-link specific scoring algorithm and a database application for storage and facilitated handling of spectra. The approach is demonstrated on LC-MS/MS experiments investigating the interaction between the recombinant constructs of human calmodulin and the actin-binding domain of plectin isoform-1a in the presence of two different cross-linkers. A comparison with competitive approaches shows Xlinq's ability to identify 33-80% complementary cross-linked peptide-spectrum-matches that are not picked up by the other methods (at 1% FDR) depending on the fragmentation method and MS2 resolution. Intra-protein cross-links identified using Xlinq are in accordance with individual structure models and inter-protein cross-link sites confirm assumed interaction interfaces at the proteins accessible surface.

References

[MCB+07] Alessio Maiolica, Davide Cittaro, Dario Borsotti, Lau Sennels, Claudio Ciferri, Cataldo Tarricone, Andrea Musacchio, and Juri Rappsilber. Structural Analysis of Multiprotein Complexes by Cross-linking, Mass Spectrometry, and Database Searching. Molecular Cellular Proteomics, 6(12):2200-2211, December 2007.

84. **Christian Peikert, Friedel Drepper, Silke Oeljeklaus, Lennart Martens and Bettina Warscheid**
PROVIS - a tool for the analysis and visualization of SILAC-based quantitative protein interaction data and beyond

Abstract: The combination of stable isotope labeling with amino acids in cell culture (SILAC) [OS02] and high resolution mass spectrometry (MS) has enabled the systematic study of changes in protein expression, protein synthesis, various posttranslational modifications, and protein-protein interactions. These studies result in large amounts of quantitative data requiring a powerful computational framework for preparation, analysis and interpretation of these data. We here present the PROVIS (PROteomic Visualization, Interaction and Statistics) system, built for storage, processing and evaluation of proteinprotein interaction data from SILAC-based affinity purification-mass spectrometry (AP-MS) experiments analyzed using MaxQuant [CJ02]. PROVIS allows for the integration of information extracted from public resources such as UniProt, Saccharomyces Genome Database (SGD), and TriTrypDB to support species-specific data mining. It also provides statistical analyses of the data to discriminate between specific protein interaction partners and co-purified contaminants. The analysis pipeline of PROVIS allows for the fast and easy generation of various modes of data presentation, which facilitates straightforward data analysis and interpretation. Moreover, for a better comprehension of co-enriched proteins, PROVIS provides different clustering approaches to determine patterns of significantly changed protein abundances across several experiments. The PROVIS system and workflow was evaluated and optimized based on experimental SILAC AP-MS data. In addition, PROVIS is currently being extended to handle different experimental setups such as siRNA- mediated protein knockdown experiments as well as experiments following label-free quantitative MS approaches.

85. **Björn Sommer, Benjamin Kormeier, Klaus Hippe, Nils Rothe, Philipp Unruh, Rudolf Warkentin and Pascal Witthus.**
CELLmicrocosmos X - Cell Modeling at the Integrative Level

Abstract: Integrative Bioinformatics combines different computer technology-related research areas to support life science research. In this work, one of the major visions of Bioinformatics is addressed: the creation of a virtual cell. The CELLmicrocosmos project combines the knowledge of ten databases with the objective to create a three-dimensional virtual cell environment integrating three cytological levels; the mesoscopic, the molecular and the functional level.

Using the CellExplorer, virtual cell environments directly or indirectly derived from microscopic data are combined with protein-related networks by using localization information acquired from different well-known web sites. In addition, these cell models can be extended by published data-based PDB membranes generated with the MembraneEditor.

86. **Astrid Wachter, Thomas Oellerich, Jasmin Corso, Ekkehard Schütz, Annalen Bleckmann, Henning Urlaub and Tim Beissbarth.**
High-Throughput Proteomic and Transcriptomic Data Integration based on MS/MS and RNA-Seq Data using Prior Pathway Knowledge

Abstract: Nowadays high-throughput technologies allow greatly accelerated research due to the high amount of information generated within a very short time. Information acquired by different technologies turn data integration into a very current topic. Although a lot of studies deal with integration of high-throughput data sets, not much emphasis is placed on incorporating prior knowledge into data integration methods yet. We implemented a three-level (protein level, transcription factor level and transcript/gene level) pathway-based integration approach which benefits from prior biological knowledge from public pathway databases. Biocarta, Reactome, KEGG and Pathway Interaction Database information was retrieved using the R package rBiopaxParser. We used this knowledge to identify pathways of differentially abundant proteins in the proteomic dataset. We reconstructed a protein network containing a selection of the most relevant pathways, identified affected transcription factors via TRANSFAC database and performed a downstream determination of corresponding target genes. This downstream analysis we compared with the upstream analysis which determined upstream transcription factors and pathways of genes corresponding to differentially expressed transcripts identified in the RNA-Seq data set. Finally, we performed an overlap analysis with the objective of evaluating the method and prior knowledge incorporation. We used this approach to integrate time-dependent MS/MS (tandem mass spectrometry) and corresponding RNA-Seq data sets generated by stimulation of human B cell receptors. With our method an extensive parallel analysis of signaling events was feasible. Of special interest for the identification of kinase/substrate networks and potential drug targets are conclusions which can be drawn in regard to the time dependency of signaling processes.

87. **Benedikt Brink, Stefan Albaum and Tim Nattkemper**
Fusion - a new polyomics platform

Abstract: High-throughput experimental technologies transformed biological research from a relatively data-poor discipline to one that is data-rich. The challenge is no longer to generate experimental data, but to compute and analyze it. A key aspect of understanding and analyzing data is visualization. A powerful visualization can make the difference whether the user is able to gain a mental model for his data and apply his biological knowledge. Furthermore, to fully understand a biological metabolism and its responses to environmental factors, it is necessary to include functional characterization and accurate quantification of all levels - gene products, proteins and metabolites, as well as their interaction.

To achieve all this, a new platform for polyomics data integration called Fusion is being developed. It shall become a center for all kinds of data-rich high-throughput experiments and offer convenient data management, powerful analysis tools, including established methods for analyzing and visualizing single omics data, as well as new features for an integrative analysis of data from multiple omics platforms. Fusion focuses on the three classical fields in omics: transcriptomics, proteomics and metabolomics and offers connections to other platforms like EMMA [DAP+09] or QuPE [ANN+09], all developed and hosted at the Center for Biotechnology (CeBiTec) in Bielefeld. It is a web based service, completely written in Java and JavaScript and is supposed to supersede ProMeTra [NAK+09].

References

[ANN+09] Stefan P. Albaum, Heiko Neuweger, Tim W. Nattkemper, Alexander Goesmann, et al. Qupe - a Rich Internet Application to take a step forward in the analysis of mass spectrometry-based quantitative proteomics experiments. Bioinformatics (Oxford, England), 25(23):3128-3134, 2009.

[DAP+09] Michael Dondrup, Stefan P. Albaum, Alfred Pühler, Alexander Goesmann, et al. EMMA 2 - a MAGE-compliant system for the collaborative analysis and integration of microarray data. BMC bioinformatics, 10:50, 2009.

[NAK+09] Heiko Neuweger, Stefan P. Albaum, Jörn Kalinowski, Alexander Goesmann, et al. Visualizing post genomics data-sets on customized pathway maps by ProMeTra - aeration-dependent gene expression and metabolism of Corynebacterium glutamicum as an example. BMC systems biology, 3:82, 2009.

88. **Marc Bonin, Sascha Johannes, Stephan Flemming, Andreas Grützkau, Florian Heyl, Irene Ziska, Pascal Schendel, Karsten Mans, Biljana Smiljanovic, Till Sörensen, Stefan Günther and Thomas Häupl.**
Development of a Database for Combined Analysis of DNA Methylation and Transcriptional Profiles in Different Immune Cells

Abstract: Background and objective: DNA methylation is one of the epigenetic mechanisms involved in the regulation of transcriptional activity and maintenance of differentiation during cell proliferation. One of the main goals was to provide an application which allows to visualise and filter high throughput data for concordant patterns of DNA-methylation and gene expression.

Materials and Methods: Transcription data were generated by Affymetrix HG-U133Plus 2.0 arrays from different immune cell types. Methylation was determined in the same cell types of the immune system using the Illumina bead array plaform. Programming was based on R, JAVA and PHP to provide a framework for analysis and storage of data.

Results: A web based application was created to determine the influence of DNA methylation. The application is designed to analyse the methylation and transcription data and present the information in a clear and easy way. Different programs and tools were combined to provide a uncomplicated workflow. To compare different populations a newly created Java-tool was applied. Data from purified immune cells provide the basis of the analysis. After upload, quality controls of the high throughput data are performed and displayed. Subsequently pairwise comparisons are calculated between array data from different populations. At this stage, it is possible to view and filter the generated data for differential expression. After analysing methylation and transcription data, both data sets can be compared by mapping CpG loci to individual genes. There are two different algorithms available to map the methylation positions and the different transcripts. The first one lists all methylation positions and transcripts, which are associated with a user specified gene. The second one finds all methylation positions and transcript which are in the range of a user specified gen and distance from the transcriptional start. Visual presentation enables the user to determine the influence of DNA methylation on the transcriptional process.

Conclusion: In summary, the new online platform provides fast and efficient processing of array data. It enables mapping of transcriptional information to genomic and epigenetic information. There was only a low number of genes which follow the general understanding of methylation induced DNA regulation. The webapplication is available under http://www.bioretis.charite-bioinformatik.de.

89. **Daniel Doerr, Jens Stoye, Sebastian Böcker and Katharina Jahn**
Algorithms for Discovering Weak Common Intervals in Indeterminate Strings

Abstract: We study indeterminate strings, which are sequences over the non-empty elements of $P(\Sigma)$, the power set of a finite alphabet Σ. Indeterminate strings have applications in computational biology where they can be used to represent chromosomes as ordered sequences of multiple co-localized genomic markers. Two intervals are called weak common intervals if each set at any position within both intervals shares least one element with any set in the other interval. We present two efficient algorithms to enumerate (1) weak common intervals and (2) approximate weak common intervals in indeterminate strings. The latter allows for a fixed number of positions where the weak common intervals condition is violated.

90. **Alexandra Kanygina, Alexander Manolov, Dmitry Alexeev and Tatyana Grigoryeva**
Proteogenomic analyses of oil sludge microbiota

Abstract: In recent years, the problem of industrial waste treatment has become essential. The recycling of oil refinery wastes is a subject of great interest because these compounds decompose very slowly and are highly toxic and thus dangerous to the environment [1]. We studied an industrial oil sludge sample in order to investigate its microbial diversity. Genome-wide analysis of P. stutzeri strain KOS6, which is the major representative of the sludge cultivated bacteria, has shown that P. stutzeri KOS6 genome contains an unusual combination of genetic traits. These include the genes associated with nitrogen fixation system (nif operon) [2] and naphthalene degradation system (nah operon). The functioning of the nah system was observed in proteomic analysis under sucrose-free conditions with naphthalene being the sole carbon and energy source. We performed a comparative genomic analysis of P. stutzeri KOS6 and 14 P. stutzeri strains available from NCBI. It has shown that the KOS6 genome contains a 122,000-bp region likely to have been acquired by horizontal gene transfer from T. auensis. This region includes phaA and phaC genes responsible for polyhydroxyalkanoates (PHA) synthesis. Finally, a phylogenetic tree of P. stutzeri strains was constructed using maximum likelihood estimation (MLE).

References

[1] Silva, Cynthia C., et al. Phylogenetic and functional diversity of metagenomic libraries of phenol degrading sludge from petroleum refinery wastewater treatment system. AMB Express, 2.1:1-13, 2012.

[2] Lalucat, Jorge, et al. Biology of Pseudomonas stutzeri. Microbiology and Molecular Biology Reviews, 70.2:510-547, 2006.