Conference proceedings German Conference on Bioinformatics (GCB) 2018

Talks

Efficient Sampling of the RNA Secondary Structure Space

Gregor Entzian [1]*, Ronny Lorenz [1], Ivo L. Hofacker [1], Andrea Tanzer [1]
(1) University of Vienna, Austria

Background
RNA secondary structures have been proven a valuable abstraction to model and understand the function of RNAs. While efficient methods exist to analyze equilibrium properties of an RNAs secondary structure ensemble, only a handful exist that model the dynamics of structure transitions. Most of them rely on exhaustive enumeration of conformations, thus restrict their application to RNAs shorter than about 100nt. Recent alternative strategies use representative local minima instead, which are derived from random samples of the ensemble. However, these methods tend to yield too many structurally similar and too few rare conformations.

Methods and Results
We developed a novel adaptive sampling strategy for RNA secondary structures to create diverse sets of structure representatives. Our algorithm iteratively computes the partition function of the secondary structure space and subsequently draws random samples according to their Boltzmann weights. For each sample, we perform a steepest descent walk and collect accessible local minima. In each round, we distort the energy function such that states receive additional weights according to their similarity to local minima over-represented in previous iterations.

Discussion
Our strategy allows us to efficiently explore the state space and even yields rare conformations that may be inaccessible to other sampling methods. A comparison against other strategies, such as non-redundant Boltzmann sampling, and Boltzmann sampling at high temperature, shows that we retrieve structurally more diverse, yet energetically important representatives. Consequently, significantly fewer representative structures are required to adequately approximate the conformation space in a way suitable for RNA folding dynamics simulations.

Structural classification of every amino acid in the human proteome

Alexander Gress [1]*, Vasily Ramensky [2], Andreas Keller [3], Olga Kalinina [1]
(1) Max-Planck-Institute for Informatics, Germany
(2) Moscow Institute of Physics and Technology, Russia
(3) Chair for Clinical Bioinformatics, Saarland University, University Hospital, Germany

Background
The mapping of genetic variations on protein three-dimensional structures can shed light of functional importance of genetic variants and is a fast developing field. Many studies performed such analysis for various datasets of known genetic variants in human. We have also recently presented a structural annotation of over 50,000 disease-associated variants (Gress et al., 2017). Structural mapping of every single possible substitution of an amino acid in the human proteome provides a background for analysis of any such a set of variant related by any common trait yet to be discovered, and gives an unprecedented insight into structural characteristics of human proteins.
Methods and results
Using our recently developed StructMAn pipeline (Gress et al., 2016), we structurally annotated all positions for all isoforms of all human proteins, in total over 37 million positions in over 93 thousand sequences. We mapped these positions in three-dimensional structures of human proteins or of proteins homologous to human and combined the structural information into a novel structural classification of variants. In doing so we increase the fraction of human proteome covered by structural annotations from 19% (all human proteins with experimentally resolved structures) to 49% (if we consider homologs).
Discussion
The results of this study can be used to perform structural annotation of any subset of genetic variants in human related to a specific trait without the usual huge computational costs. Annotation of genetic variants of single individuals can be done in real time, which can be important for personalized medicine.

Discovering conserved motifs in protein three-dimensional structures using frequent subgraph mining

Sebastian Keller [(1)]*, Pauli Miettinen [(1)], Olga Kalinina [(1)]
(1) Max Planck Institute for Informatics, Germany

Background
Conserved protein three-dimensional (3D) motifs, often corresponding to functionally important residues, present a considerable challenge for motif-detection tools, since they can occur in distantly related proteins that cannot be aligned based on their sequence similarity. Here we present RINminer, a graph-based approach to identify conserved structural motifs that relies on residue interaction networks (RINs) of protein 3D structures.
Methods and results
We base RINminer on gSpan, an established frequent subgraph mining algorithm. Given a set of input graphs corresponding to RINs, where each node represents a residue and each edge an interaction between them, we determine all subgraphs that are shared by at least a user-defined number of graphs. As novel features, RINminer supports multidimensional edge labels and, most importantly, allows deviations of labels corresponding to the distances between non-covalently interacting residues. We also use these deviations to rank the subgraphs by their structural conservation and to reduce the search space of the algorithm. We validated our approach using 3D structures of 120 protein families that share a common PROSITE sequence pattern, achieving on average an overlap of 51.4% between the highest-scoring subgraphs and the patterns' strictly conserved residues.
Finally, we applied our approach to 543 SCOP superfamilies, identifying many previously

described motifs along with novel structurally conserved ones.
Discussion
The presented method demonstrates that protein 3D structures can be effectively mined using graph-based techniques. Additionally, it has important applications for the discovery of protein functional motifs and potential extensions for mining protein-protein and protein-ligand interactions.

Practical approaches for exploring structural variants in whole-genome sequencing data

Birte Kehr [(1)]*
(1) Berlin Institute of Health, Germany

Background: Genetic variants influence susceptibility to disease and response to medical treatment. Comprehensive variant identification is continuously fostered by on-going advancements in sequencing technologies. These advancements include the increase of throughput in short-read sequencing as well as the introduction of new protocols and technologies such as linked-read and long-read sequencing. However, tailored computational approaches are required for each technology that account for the specifics of each data type when searching for variants that make up an individual.
Methods: The current research focus of the Genome Informatics junior research group at the Berlin Institute of Health lies on algorithm development for the detection and genotyping of structural variants (SVs) from various types of whole-genome sequencing data. We leverage the increasing throughput of short-read sequencing technologies and develop scalable approaches that can discover SVs in tens of thousands of individuals simultaneously, harnessing the joint analysis of multiple individuals for improved recall and precision. Furthermore, we develop methods for the analysis of new data types: long reads with high error rates and linked reads with long-range information in barcodes that can resolve SVs involving repeats.
Discussion: Our research is driven by the characteristics of the data and biomedical needs. Application of our implementations in computationally efficient and user-friendly tools has led to the discovery of variants of interest for numerous diseases. Thus, by developing new computational tools we strive to deepen our understanding of genetic variation and its role in disease.

Modelling tumour evolution from single-cell sequencing data

Katharina Jahn [(1)]*, Jack Kuipers [(2)], Niko Beerenwinkel [(1)]
(1) ETH Zurich, Switzerland
(2) ETH Zurich, D-BSSE, Computational Biology Group, Switzerland

Background: Through recent advances in sequencing technology it is now possible to study the mutational history of cancer development at the level of single cells. The goals of this in-depth analysis are to tailor cancer treatments to individual patients based on the genetic composition of their tumour and thereby increase treatment efficacy and reduce side effects. The analysis of single-cell sequencing data poses a number of statistical challenges such as

elevated noise rates due to allelic drop out, missing data, uneven coverage and contamination with doublet samples.
Methods and results: We developed a Bayesian inference scheme for tumour mutation histories based on single-cell sequencing data [Jahn et al., 2016]. In this talk I will focus on two recent extensions of this work, a novel single-cell mutation caller that takes the underlying cell phylogeny into account [Singer et al., 2018] and a rigorous statistical test to identify the presence of parallel mutations and mutational loss [Kuipers et al., 2017]. Our results on simulated and real tumour data show that a thorough modelling of the noise inherent to single-cell data allows for an accurate reconstruction of tumour mutation histories.
Discussion: While mutation histories inferred from single-cell DNA data exhibit unprecedented resolution, the underlying tree models including our own are currently limited in that they focus exclusively on SNVs. The integration of copy number and structural variants is largely an open problem.
[Jahn et al., 2016] Jahn, K., Kuipers, J., and Beerenwinkel, N., 2016. Tree inference for single-cell data. Genome Biology, 17:86.
[Kuipers et al., 2017] Kuipers, J., Jahn, K., Raphael, B. & Beerenwinkel, N., Genome Research 27 (11), 1885-1894.
[Singer et al., 2018] Singer, J. , Kuipers, J, Jahn, K. & Beerenwinkel, N., Under review.

Modeling tumor progression using cyclic interaction networks

Rudolf Schill [(1)]*, Rainer Spang [(1)], Tilo Wettig [(1)], Stefan Solbrig [(1)]
(1) University of Regensburg, Germany

Background: Tumors turn malignant by accumulating genetic alteration events which tend to fixate in specific chronological sequences. This is due to unknown interactions between events which we aim to infer from cross-sectional data, consisting of co-occurence patterns in different tumors at different stages of development.
So far, increasingly general types of models have been proposed: events that facilitate each other in linear chains (Vogelstein, 1988), trees (Desper, 1999), directed acyclic graphs (Beerenwinkel, 2007) and cyclic networks (Hjelm, 2006). However, they cannot directly account for patterns of mutual exclusivity, which are common for genes that participate in the same regulatory or signaling pathway. Current approaches explicitly model pathways as non-overlapping groups of mutually exclusive events (Raphael, 2015; Cristea, 2017).
Methods and results: We propose Mutual Hazard Networks (MHN) which characterize events by their spontaneous rate of fixation, and by multiplicative effects they exert on the rates of successive events. These effects can be cyclic and greater or less than one, i.e., facilitating or inhibiting.
The n^2 parameters are inferred by treating the MHN as an acyclic Markov process on a 2^n dimensional space state space. Its rate matrix is decomposed into a sum of Kronecker products for computational efficiency.
Our method outperforms earlier models in cross-validated model fit and can account for overlapping pathways.
Discussion: We have introduced a new framework for modeling tumor progression which naturally extends earlier Bayesian and cyclic networks. It shows promising results, but is so far limited for systems with up to 20 - 25 events.

Finding long-range RNA-RNA-Interaction with third generation sequencing techniques

Kevin Lamkiewicz [(1)]*, Manja Marz [(1)]
(1) Friedrich Schiller University Jena, Germany

Background
With the rise of third generation sequencing technologies (TGS) reads from sequencing experiments dramatically increase in length. One of the many potentials of long-read data is the prediction of long-range RNA-RNA-interactions (LRIs), which is a commonly used mechanism in viruses to initialize replication.
With conventional NGS experiments co-occuring compensatory mutations had to be inferred from short-reads, making precise predictions difficult and inaccurate.
However, with long-read data we are able to scan one single read spanning several thousands of nucleotides for potential LRIs.
Methods and results
Here we present LOCATION, our pipeline that scans mapped long reads for nucleotide pairs that may be part of an LRI. A pair of nucleotides is considered as a candidate, if it passes several filter steps, like coverage, overall and compensatory mutation rate and the base-pairing probability on RNA secondary structure level. Using simulated data of an Hepatitis C virus genome, we are able to correctly distinguish between introduced compensatory mutations and sequencing errors. We provide an interactive plot visualizing all candidates in a comprehensive way.
Discussion
As TGS is on the rise, more and more data will be available. We offer LOCATION as a ready-to-use pipeline that can predict LRIs from sequencing data. Results on artificial data show that the high error rates of sequencers do not produce false positive candidates. Thus we are confident that LOCATION can be used on real viral TGS data which may be available in the future.

vRNAsite: Prediction and evaluation of viral RNA-RNA interaction sites between influenza A wild type and mutant vRNAs

Daniel Desirò [(1)]*, Hardin Bolte [(2)], Martin Schwemmle [(2)], Manja Marz [(1)]
(1) Friedrich Schiller University Jena, Germany
(2) Albert Ludwig University of Freiburg, Germany

Background
Seasonal influenza A virus (IAV) epidemics are difficult to counter with vaccination. This is due to the unique packaging of the eight viral RNA (vRNA) segments and the ability to create reassortants between different strains. One theory about packaging involves vRNA-vRNA interactions (RRI) between different segments [1,2], however to our knowledge there has been no attempt to computationally analyze the involved mechanisms. Here, we present vRNAsite which predicts possible RRIs and evaluates them to determine potential interaction networks between all vRNAs.
Methods and Results
vRNAsite takes the original and mutated segments as input and a sliding window between

any two vRNAs calculates minimum free energy (MFE) dependent scores for each single nucleotide pairing via RNAcofold [3]. Resulting scores are used to evaluate possible interaction sites as well as differences between WT and mutant RRIs. Chained interactions are then computed with a directed acyclic graph, where each vertex represents a different set of vRNA segments connected by weighted edges based on the RRI scores. The maximum scoring path from all starting single segment vertices to the vertex containing all eight segments is then reported as a possible packaging network.

Discussion

While one of the predicted RRIs has been validated in vitro [2], the majority of them still have to be verified to assess the accuracy of the potential packaging networks. Currently the scoring is based on MFE only, but it would be preferable to also include SHAPE-MaP [4] data to improve the predictions. This might yield the capability to predict the pathogenic potential and reproductive capacity of reassortant IAV.

References

[1] Fournier E, et al. (2012). A supramolecular assembly formed by influenza a virus genomic RNA segments. Nucleic Acids Res., 40:2197–2209.

[2] Gavazzi C, et al. (2013). A functional sequence-specific interaction between influenza a virus genomic RNA segments. Proc. Natl. Acad. Sci. U.S.A., 110:16604–16609.

[3] Lorenz R, et al. (2011). ViennaRNA package 2.0. Algorithms for molecular biology : AMB, 6:26.

[4] Siegfried N A, et al. (2014). RNA motif discovery by SHAPE and mutational profiling SHAPE-MaP. Nature Methods, 11(9):959.

Analysis of RNA-RNA interaction data

Richard A. Schäfer [(1)]*, Björn Voß [(1)]

(1) Computational Biology, Institute for Biochemical Engineering, University of Stuttgart, Germany

Introduction

RNA-RNA inter- and intramolecular interactions are fundamental for multiple biological processes. While there are reasonable approaches to map RNA secondary structures genome-wide, understanding how different RNAs inter- act to carry out their regulatory functions requires mapping of intermolecular base pairs. Recently, different strategies to detect RNA-RNA duplexes in cells, termed Direct Duplex Detection (DDD) methods, have been developed. Common to all is that they rely on Psoralen mediated in vivo crosslinking and RNA Proximity Ligation (RPL) joining the two interacting RNA strands. Subsequently, the RNA is sequenced via RNA-seq and analysed with respect to inter- and intramolecular RNA-RNA interactions. In this work, we present a general automated pipeline for the inference of RNA-RNA interactions from raw DDD reads.

Methods and Results

We applied our pipeline to data from different Direct Duplex Detection methods and compared our results to the original ones and show an improvement in the detection of RNA-RNA interactions within the datasets. Our approach starts with the preprocessing of the reads, in which the raw single- or paired-end reads are filtered and adapter-trimmed. Afterwards, the primary data analysis starts with the alignment of the reads to the reference which is followed by the split read calling. In particular, applying tolerant alignment settings, a multitude of split

reads can be detected for further analysis. With the aid of a statistical approach, the calculation of different metrices (e.g., complementarity, energy minimization) and clustering of the split reads data, RNA-RNA interactions can be reliably assessed. In this regard, our implementation shows significant improvements in performance as opposed to custom pipelines used in the original publications.
Discussion
This showed that our method, due to its tolerant primary data analysis reconstructs more information about known and novel RNA-RNA interactions that otherwise would have been lost. Our software can also be applied to datasets from related experimental methods, such as CLASH or RIL-seq. Thus, we present the first fully integrated tool for the analysis of RNA-RNA interaction data. It will foster the elucidation of RNA structure and RNA-based regulation.

Watchdog – a workflow management system for the distributed analysis of large-scale experimental data

Michael Kluge [(1)], Caroline C. Friedel [(1)]*
(1) Ludwig Maximilian University of Munich, Germany

Background:
The development of high-throughput experimental technologies, such as next-generation sequencing, have led to new challenges for handling, analyzing and integrating the resulting large and diverse datasets. Bioinformatical analysis of these data commonly requires a number of mutually dependent steps applied to numerous samples for multiple conditions and replicates. To support these analyses, a number of workflow management systems (WMSs) have been developed to allow automated execution of corresponding analysis workflows. Major advantages of WMSs are the easy reproducibility of results as well as the reusability of workflows or their components.
Methods & Results
In this article, we present Watchdog, a WMS for the automated analysis of large-scale experimental data. Main features include straightforward processing of replicate data, support for distributed computer systems, customizable error detection and manual intervention into workflow execution. Watchdog is implemented in Java and thus platform-independent and allows easy sharing of workflows and corresponding program modules. It provides a GUI for workflow construction and a helper script for creating new module definitions. Execution of workflows is possible using either the GUI or a command-line interface and a web-interface is provided for monitoring the execution status and intervening in case of errors.
Discussion:
Watchdog is a powerful and flexible WMS for the analysis of large-scale high-throughput experiments. We believe it will greatly benefit both users with and without programming skills who want to develop and apply bioinformatical workflows with reasonable overhead. The software, example workflows and a comprehensive documentation are freely available at www.bio.ifi.lmu.de/watchdog.

Massively parallel and scalable enumeration of minimal pathways in metabolic networks

Bianca Allegra Buchner [(1)]*, Jürgen Zanghellini [(1)]
(1) acib GmbH (Austrian Centre of Industrial Biotechnology), Austria

Background: Elementary flux mode (EFM) analysis is a commonly used method for the unbiased characterization of cellular metabolism. EFMs are unique and minimal steady-state pathways that represent the smallest functional units in metabolism. The full set of EFMs characterizes the totality of an organism's metabolic capabilities. Mathematically, the enumeration of EFMs is equivalent to vertex enumeration in convex polytopes. State-of-the-art software solutions are all based on variants of the (Fourier-Motzkin) double description method. However, with current methods the enumeration of EFMs in genome-scale metabolic models is computationally intractable as the memory demand scales combinatorially with the size of the metabolic network.
Methods and results: We use lexicographic reverse search (lrs) for the enumeration of EFMs in large-scale metabolic networks. Making use of a highly parallelized lrs-implementation and extensive lossless (network) compression, we show that the EFM enumeration in metabolic networks is almost embarrassingly parallel and strongly scalable. In fact, EFMs in medium-scale metabolic networks can be quickly enumerated already with medium-sized computer cluster with negligible memory requirements. Although the complete enumeration of EFMs in genome-scale metabolic models remains out of reach, we show that at least the set of all (yield-)optimal EFMs can be enumerated. Thus, for the first time an unbiased analysis of alternate optima in flux-balance applications becomes possible in actual research practice.
Discussion: Traditionally lrs has been used for non-degenerated polytopes. Metabolic networks, however, are highly degenerated and lrs-based methods are considered to be unsuitable. Here we demonstrate that this assertion does no longer hold.

Characterization and design of phase spaces and yield spaces in genome-scale metabolic models

Steffen Klamt [(1)]*, Stefan Müller [(2)], Georg Regensburger [(3)], Jürgen Zanghellini [(4)]
(1) Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
(2) Faculty of Mathematics, University of Vienna, Austria
(3) Institute for Algebra, Johannes Kepler University Linz, Austria
(4) Austrian Centre of Industrial Biotechology, University of Natural Resources and Life Sciences, Vienna, Austria

Background: Production rates and yields are key parameters of biochemical transformation processes. While optimal rates and phase spaces are readily studied with flux-balance analysis (FBA) approaches, optimal yields and yield spaces are rarely systematically analyzed. Often elementary flux modes (EFMs) are used to characterize yields and yield-optimal pathways. However, EFMs characterize the unbounded flux cone and are incompatible with non-zero flux bounds and allocation constraints often used in FBA.
Methods and results: By resorting to the concept of elementary flux vectors (EFVs), it is possible to generalize the idea of unique metabolic pathways to also account for inhomogeneous linear constraints. We show that any rate-optimal FBA solution sits in an optimal polyhedron spanned by (certain) EFVs. This holds true not only for rate-optimal but

for yield-optimal solutions too, which cannot be found by standard FBA approaches. Next, we demonstrate that (optimal) yield spaces can be readily calculated even in genome-scale metabolic models by linear-fractional programing without explicitly enumerating EFVs. Although phase spaces and yield spaces often are of similar shapes (and therefore sometimes confused), they carry very different information. In a realistic analysis based on E. coli, we show how these complementary pieces of information can be used to understand and optimally shape the metabolic capabilities of cell factories with any desired yield and/or rate requirements.
Discussion: We conclude that EFVs provide an unifying framework for the theoretical description and analysis of any constraint-based model under arbitrary linear constraints. More specifically, EFVs close the gap between biased FBA approaches and unbiased EFM approaches and allow one to fully characterize and shape metabolic phase spaces and yield spaces. This reinforces the fundamental importance of EFVs (or EFMs) as the ``coordinates of metabolism´´. However, an explicit enumeration of EFVs is not required as phase spaces and yield spaces can be efficiently computed even in genome-scale metabolic networks.

Integration of Transcription Factor Binding Data with scRNA-Seq

Mahmoud Ibrahim [(1)]*
(1) RWTH Aachen University, Germany

Introduction
Single-cell RNA-Seq (scRNA-Seq) data can revolutionize cell type identification from complex tissues and organs. While cell clustering from scRNA-Seq data provides valuable information, the results are often difficult to interpret due to the sparsity of the data and the lack of a functional understanding of the regulatory mechanisms underlying cell variability.

Results
I describe a simple algorithm for integrating scRNA-seq with bulk or single-cell transcription factor binding information (ChIP-Seq or ATAC-seq). The algorithm co-clusters cells and transcription factors, annotating each cell type with a distinct transcription factor regulatory program. Relying on the well-known Singular Value Decomposition, the algorithm is fast, easy to understand and can be easily applied to scRNA-Seq data from any platform including DROP-Seq variants assaying thousands of cells. The method is extended further to a clustering routine where genes are clustered in low-dimensional space while iteratively optimizing cell and transcription factor clusters, also allowing for identification of cell-type markers. I demonstrate the algorithm using simulated data and published single-cell datasets. Compared to state-of-the-art cell clustering methods, this method can accurately delineate cell clusters while elucidating their latent regulatory mechanisms, providing much richer information.

Discussion
Although it was developed with integrating scRNA-Seq and transcription factor binding data in mind, the methods described here are in fact applicable to any two highly-dimensional datasets measured on the same gene set. This work demonstrates that integrative methods combining scRNA-Seq data with complimentary data can increase the power of cell clustering analyses and provide a clearer understanding of expression variability.

Integrative prediction of gene expression with chromatin accessibility and conformation data

Florian Schmidt [(1)], Fabian Kern [(1)]*, Marcel Schulz [(1)]
(1) Cluster of Excellence on Multimodal Computing and Interaction, Saarland Informatics Campus, Germany

An important task in Bioinformatics is elucidating transcriptional regulation through transcription factors (TFs). Several studies have shown that TFs do not exclusively bind to promoter regions of genes but also to distally located enhancer regions that loop to gene regions in 3D space, leading to potentially large genomic distances between enhancer and gene regions.
We compared several means to assign distal regulatory regions identified via TF-ChIP-seq, DNase1-seq, and Hi-C experiments to individual genes. In detail, we assessed the quality of prevalent window-based and nearest-gene associations using per-sample gene expression prediction. Our results suggested that the widely used nearest-gene assignment is outperformed by window-based approaches.
Furthermore, we attempted to improve the window-based predictions by extending our TEPIC framework to incorporate data from chromatin conformation capture experiments, e.g. Hi-C, to link distal enhancers to their putative target genes.
We found that including Hi-C data (best resolution 5kb) did not lead to an improvement in model performance, because too many false positives, e.g. TF-ChIP-seq regions, were included in the large Hi-C regions.
Both nearest-gene and window-based assignments do not generalize well over all possible regulatory landscapes, which can be explained by the existence of gene dense and gene sparse regions. In addition, our results indicated that current Hi-C experiments have an insufficient resolution to establish accurate genome-wide promoter-enhancer interactions. However, we showed that incorporating chromatin state segmentations by ChromHMM effectively reduces false positive assignments of regulatory regions to their presumed target genes in several data setups tested.

Integrated whole-genome and targeted DNA methylation analysis with EPIC-TABSAT

Julie Krainer [(1)]*, Manuela Hofner [(1)], Walter Pulverer [(1)], Andreas Weinhäusel [(1)], Klemens Vierlinger [(1)], Stephan Pabinger [(1)]
(1) AIT - Austrian Institute of Technology, Center for Health and Bioresources, Molecular Diagnostics, Vienna, Austria

Background
DNA methylation plays an important regulatory effect on gene expression and is a suitable biomarker for disease diagnosis and treatment response prediction. Samples can be analyzed by whole-genome bisulfite sequencing, RRBS, or epigenome-wide methylation profiling microarrays (Illumina EPIC chips) to identify methylation aberrances. To translate these findings into an application or a clinical setting, targeted deep bisulfite sequencing (TDBS) has emerged as a flexible and cost-effective method, especially attractive for

validation studies. As these studies need to be applied in a standardized setting, a robust, user-friendly, yet flexible software is required.
Methods and results
Here we present EPIC-TABSAT, an easy-to-use tool for the analysis of targeted bisulfite sequencing data. In addition to covering the complete data analysis workflow from quality assessment, methylation calling to interactive result presentation, it supports the combined analysis of TDBS with whole-genome EPIC data. Especially designed for bisulfite resequencing studies, the interactive visualization of targets allows the investigation of the methylation state of every CpG for both EPIC and TDBS data. Each CpG site is enriched with meta information such as restriction cut sites or gene models. Moreover, sample specific methylation patterns are calculated and visualized.
Discussion
EPIC-TABSAT bridges epigenome-wide studies with medium-scaled investigations and is suitable for use in clinical and validation settings. The presented version contains major improvements to the freely available tool at http://github.com/tadkeys/tabsat. It provides a new way to study methylation patterns, and offers an unprecedented way to analyze and interpret TDBS data in combination with epigenome wide methylation studies.

Correctly counting molecules using unique molecular identifiers

Florian Pflug [(1)]*, Arndt von Haeseler [(1)]
(1) Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories (MFPL), Austria

Background:
Counting molecules using next-generation sequencing (NGS) suffers from PCR amplification bias, which reduces the accuracy of many quantitative NGS-based experimental methods such as RNA-Seq. This is true even if molecules are made distinguishable using unique molecular identifiers (UMIs) before PCR amplification, and distinct UMIs are counted instead of reads: Molecules that are lost entirely during the sequencing process will still cause underestimation of the molecule count, and amplification artifacts like PCR chimeras create phantom UMIs and thus cause over-estimation.
Methods & Results:
We introduce the TRUmiCount algorithm to correct for both types of errors. The TRUmiCount algorithm is based on a mechanistic model of PCR amplification and sequencing, whose two parameters have an immediate physical interpretation as PCR efficiency and sequencing depth and can be estimated from experimental data without requiring calibration experiments or spike-ins. TRUmiCount captures the main stochastic properties of amplification and sequencing, and allows us to filter out phantom UMIs and to estimate the number of molecules lost during the sequencing process. Simulations show that the phantom-filtered and loss-corrected molecule counts computed by TRUmiCount measure the true number of molecules with considerably higher accuracy than the raw number of distinct UMIs.
Discussion:
TRUmiCount increases the accuracy of transcript counts over simply counting unique UMIs for a wide range of sequencing depths, including depths that a typical for single-cell RNA-seq experiments. Its PCR efficiency estimate can provide insight into the amplification step, and can be used to detect low-quality libraries or cells in single-cell RNA-seq.

We recently published a paper about TRUmiCount in Bioinformatics, see https://doi.org/10.1093/bioinformatics/bty283

PhyloProfile: an interactive visualization tool for exploring enriched phylogenetic profiles

Ngoc-Vinh Tran [(1)]*, Carla Mölbert [(1)], Ingo Ebersberger [(2)]
(1) Goethe University Frankfurt, Germany
(2) Goethe University Frankfurt, Senckenberg Biodiversity and Climate Research Centre, Germany

Background
Phylogenetic profiles form the basis for tracing proteins and their functions across species and through time. Nowadays, novel genome sequences emerge on a daily basis, and often represent species from the remotest corner of the tree of life. A dynamic and interactive analysis even of large phylogenetic profiles is, thus, essential for functionally annotating this data and for integrating it into a comprehensive picture of organismal evolution.
Methods and Results
Here, we present PhyloProfile (https://github.com/BIONF/phyloprofile), a shiny-R based tool for interactively viewing and exploring phylogenetic profiles that can be enriched with up to two additional information layers, e.g. the extent to which two proteins agree in their domain architecture. PhyloProfile allows dynamically adapting the resolution of an analysis from an overview across hundreds of proteins and hundreds of taxa to the pair-wise comparison of individual proteins at the level of their domain architectures without the need to modify input data. The tool provides several analysis functions, such as (i) identifying related proteins in the same metabolic pathways, (ii) estimating gene ages, (iii) determining core protein sets, or (iv) searching for proteins with a lineage specific change in their domain architectures.
Discussion
A meaningful in silico functional annotation transfer from model- to non-model organisms optimally should integrate homology information with additional evidences that support a functional equivalence of two proteins. PhyloProfile closes a methodological gap as it facilitates an intuitive way to screen and curate even large phylogenetic profiles in the quest for functionally similar – or diverged – proteins.

Tracing functional protein interaction networks using a feature-aware phyletic profiling

Holger Bergmann [(1)]*, Julian Dosch [(1)], Ingo Ebersberger [(2)]
(1) Goethe University Frankfurt, Germany
(2) Goethe University Frankfurt; Senckenberg Biodiversity and Climate Research Centre (BIK-F), Frankfurt, Germany

Introduction
Tracing the phyletic distribution of protein interaction networks across hundreds to thousands of species calls for scalable and reliable methods for assessing evolutionary relationships and functional similarity of proteins. Standard ortholog search tools scale quadratic with the number of taxa and have prohibiting running times. Moreover, the resulting presence-absence

patterns of orthologs across taxa provide no information about their functional similarity or divergence.
Methods and Results
HaMStR-OneSeq (https://github.com/BIONF/HaMStR) integrates for a seed protein a targeted ortholog search with an assessment of the pair-wise feature architecture similarity (FAS) between the seed and its orthologs. For the ortholog search, HaMStR-OneSeq applies a hidden Markov model (HMM) based approach, which scales linearly with the number of search taxa. The HMM training data is iteratively compiled on-the-fly from species with increasing phylogenetic distance from the seed taxon. For the FAS scoring, HaMStR-OneSeq considers identity, copy number, and positional similarity of shared features, e.g. Pfam domains, between two proteins. The feature architecture is implemented as a directed acyclic graph. If redundant features overlap in an architecture, we identify the linearized path maximizing the FAS score. Feature-aware phylogenetic profiles of HaMStR-OneSeq can be visualized, interactively explored, and analyzed in PhyloProfile (https://github.com/BIONF/PhyloProfile).
Discussion
HaMStR-OneSeq facilitates the dynamic generation of feature-aware phylogenetic profiles for a set of seed proteins across large and customizable taxon collections. The simultaneous assessment of both presence/absence of orthologs across species, and of their similarities/deviations in the feature architecture eases the tracing of proteins and their function across species and through time.

High genomic diversity of multi-drug resistant wastewater Escherichia coli

Norhan Mahfouz [(1)]*, Serena Caucci [(2)], Eric Achatz [(3)], Torsten Semmler [(4)], Sebastian Guenther [(5)], Thomas Berendonk [(3)], Michael Schroeder [(3)]
(1) Technical University in Dresden - TUD, Austria
(2) UNU-FLORES, Germany
(3) Technical University in Dresden - TUD, Germany
(4) Robert Koch Institute - FUB, Germany
(5) University of Greifswald, Germany

Wastewater treatment plants play an important role in the emergence of antibiotic resistance. They provide a hot spot for exchange of resistance within and between species. Here, we analyse and quantify the genomic diversity of the indicator Escherichia coli in a German wastewater treatment plant and we relate it to isolates' antibiotic resistance. Our results show a surprisingly large pan-genome, which mirrors how rich an environment a treatment plant is. We link the genomic analysis to a phenotypic resistance screen and pinpoint genomic hot spots, which correlate with a resistance phenotype. Besides well-known resistance genes, this forward genomics approach generates many novel genes, which correlated with resistance and which are partly completely unknown. A surprising overall finding of our analyses is that we do not see any difference in resistance and pan genome size between isolates taken from the inflow of the treatment plant and from the outflow. This means that while treatment plants reduce the amount of bacteria released into the environment, they do not reduce the potential for antibiotic resistance of these bacteria.

HOPS: A pipeline for screening archaeological remains for pathogen DNA

Ron Huebler [(1)]*, Felix M Key [(1)], Christina Warinner [(1)], Kirsten Bos [(1)], Johannes Krause [(1)], Alexander Herbig [(1)]
(1) Max Planck Institute for the Science of Human History, Germany

Background
Large-scale metagenomic studies conducted on the molecular data from archaeological remains can provide further insights into host-pathogen relationships throughout human history.
Here we present HOPS (Heuristic Operations for Pathogen Screening), an automated pathogen screening pipeline for ancient DNA sequence data that provides straightforward and reproducible information on species identification and authentication of their ancient origin. HOPS consists of (1) a version of MALT, (2) MaltExtract, a Java tool that evaluates aDNA authenticity criteria for a list of target species, and (3) customizable post-processing scripts to identify candidate hits.
Methods and results
We evaluated HOPS with DNA sequences obtained from archaeological samples and simulated aDNA data created by spiking 33 bacterial pathogens of interest into diverse metagenomic backgrounds. We further tested HOPS on five control samples negative for pathogens. We could further use these data to test and adjust for biases generated from the database contents and structure. Finally, we compared the performance of HOPS to two other methodologies for microbiome characterization: a maker gene based approach with MIDAS and k-mer matching with Kraken.
Discussion
HOPS successfully identified all target pathogens in simulation and the sample data.
For simulated sets with only a miniscule amount of DNA (50 Reads) HOPS was more sensitive than MIDAS and Kraken and was able to avoid false positive detections in the negative control samples. Overall, HOPS provides a versatile and fast pipeline for high-throughput pathogen screening of archaeological material that aids in the identification of candidate samples for further analysis.

Pan-genome mapping and pairwise SNP-distance improve detection of Mycobacterium tuberculosis transmission clusters

Christine Jandrasits [(1)]*, Stefan Kröger [(1)], Walter Haas [(1)], Bernhard Renard [(1)]
(1) Robert Koch-Institut, Germany

Background:
With about 1.7 Million deaths and over 10 Million new infections per year Tuberculosis is a major threat to global health. It is essential to detect and interrupt transmissions to stop the spread of this disease. With the rising use of next-generation sequencing, its application in the surveillance of Mycobacterium tuberculosis has become increasingly popular. The main goal of molecular surveillance is the identification of patient-patient transmission and transmission clusters. This can be supported by measuring the distance between isolates based on single nucleotide polymorphisms.

The mutation rate of M. tuberculosis of 0.3–0.5 mutations per genome per year is very low. Many existing methods for comparative analysis of isolates provide inadequate results for such stable genomes since their resolution is too limited.
Methods and results:
We developed a new approach for comparing pairs of isolates that takes into account every detectable difference for each pair. As the choice of the reference sequence strongly impacts variant detection results, we combine our improved SNP-distance calculation with the use of a pan-genome, incorporating the genomic information of more than 100 M. tuberculosis reference genomes.
We compare our pairwise method with previously published approaches for SNP-distance measurement. We classify relations between over 300 samples in a simulated set with an accuracy of 0.99 and show the improvement of transmission cluster detection in real datasets.
Discussion:
We demonstrate that using a pan-genomic reference and pairwise SNP-distances improves the collective analysis and comparison of large groups of similar and diverse M. tuberculosis isolates.

Improving homology-based gene prediction using intron position conservation and RNA-seq data

Jens Keilwagen [(1)]*, Frank Hartung [(1)], Michael Paulini [(2)], Sven O. Twardziok [(3)], Jan Grau [(4)]
(1) Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Germany
(2) EMBL-EBI, United Kingdom
(3) Plant Genome and Systems Biology, Helmholtz Center Munich - German Research Center for Environmental Health, Germany
(4) Institute of Computer Science, Martin Luther University Halle-Wittenberg, Germany

Background: Next Generation Sequencing technologies enable rapid and cost-efficient sequencing of genomes. However, after sequencing and assembling the genome of an organism, it is important to provide annotations, especially of protein-coding genes.
Methods and results: Here, we present a substantially extended and improved version of the homology-based gene prediction program GeMoMa (Keilwagen et al., BMC Bioinformatics, 2018). The new version of GeMoMa predicts protein-coding genes based on existing gene models of related species utilizing amino acid sequence conservation, intron position conservation, and additional experimental evidence from RNA-seq data. In addition, GeMoMa now allows for including multiple reference species yielding improved sensitivity and/or specificity of gene predictions.
We show on published benchmark data for plants, animals and fungi that GeMoMa performs better than the gene prediction programs BRAKER1, MAKER2, and CodingQuarry, and purely RNA-seq-based pipelines for transcript identification. In addition, we demonstrate that using multiple reference organisms may help to further improve the performance of GeMoMa. Finally, we apply GeMoMa to four nematode species and to the recently published barley reference genome indicating that current annotations of protein-coding genes may be refined using GeMoMa predictions.
Discussion: GeMoMa might be of great interest for researchers annotating newly sequenced genomes, for genome curators scrutinizing existing gene annotations, as well as for researchers interested in a specific gene or gene family, looking for homologs in another

species. Hence, we make GeMoMa available under GNU GPL3 at http://www.jstacs.de/index.php/GeMoMa including a user manual and documentation.

Fast and Accurate Distance Computation from Unaligned Genomes

Fabian Klötzl [(1)]*, Bernhard Haubold [(2)]
(1) MPI for Evolutionary Biology, Germany
(2) Max-Planck-Institute for Evolutionary Biology, Germany

Background
Traditionally, methods for phylogeny reconstruction require an alignment. However, for large samples of whole genomes alignment computation becomes unfeasible. Thus, in recent years research on “alignment free” methods has intensified. Some of these methods are very fast, usually accompanied by some loss of accuracy. For example, “mash” (Ondov et al., 2016) uses hashing to rapidly compute useful genome distances. However, distances between closely related sequences, as obtained by tracking pathogen outbreaks, are overestimated. On the other hand, spaced-words methods solve the accuracy problem at the expense of time. We propose a new method based on approximate local alignments that is faster than spaced-words and more accurate than hashing.
Methods & Results
Our new tool “phylonium” picks one reference from a given sequence sample and computes approximate alignments of all other sequences against this reference. We then assume transitivity of homology and compare all sequences which aligned to the same segment on the reference. From this the evolutionary distances between the sequences are estimated. We tested “phylonium” on a number of simulated and real-world data sets. For example, on a set of 109 Escherichia coli genomes “phylonium” correctly identifies the main clades. Furthermore, the estimated branch lengths are comparable to other state-of-the-art tools, while “phylonium” is computationally less demanding.
Discussion
“Phylonium” is more accurate than hash-based methods of distance computation, and faster than spaced-words methods. We are currently exploring where exactly on the accuracy/speed continuum “phylonium” is located.

Computing Genome Mappability Scores for Faster Read Mapping using Optimum Search Schemes

Christopher Pockrandt [(1)]*, Kiavash Kianfar [(2)], Bahman Torkamandi [(2)], Haochen Luo [(2)], Knut Reinert [(3)]
(1) Max Planck Institute of Molecular Genetics / Free University of Berlin, Germany
(2) Texas A&M University, United States
(3) FU Berlin, Germany

Background
==========
As approximate string matching is not only one of the most fundamental problems in

Bioinformatics but also a limiting factor in terms of run time in many applications such as read mapping, a lot of research is being dedicated to this area. Approximate mapping of reads from repetitive regions is the computationally most expensive part in read mapping.
One is usually only interested in the mapping positions of rather unique parts of the genome. To address this issue we currently develop approximate string matching algorithms in indices that take the uniqueness (referred to as mappability) of all positions in the genome into account.
In this talk we are taking up on the work by Derrien et al. [1] and show how to compute the Genome Mappability Score a magnitude faster by using the recently published Optimum Search Schemes [2].
Methods and results
==========
Given a read length K and a maximum number of errors E, the number of occurrences of each K-mer in the genome with up to E errors is computed. Since computing mappability scores can easily take up to several hours or even days depending on K and E, we developed an approach using Optimum Search Schemes (a search strategy that avoids redundant search steps in an index) that additionally avoids duplicate computations of overlapping K-mers. Furthermore we use a hybrid approach of indexed search and in-text verification.
Discussion
==========
Being able to quickly compute the mappability scores of an entire genome, even for up to 3 errors makes it feasbile to not only incorporate mappability scores into post-mapping analysis and interpretation but also include it into the mapping process and speed up string matching for read mapping to repetitve regions.
[1] Derrien T, et al. (2012) Fast Computation and Applications of Genome Mappability. PLoS ONE 7(1): e30377.
[2] Kianfar K., et al. (2018) Optimum Search Schemes for Approximate String Matching Using Bidirectional FM-Index. RECOMB-Seq.

Integrating Time-Series Data with Network Enrichment

Tim Kacprowski [(1)]*, Christian Wiwie [(2)], Jan Baumbach [(1)]
(1) Chair of Experimental Bioinformatics, Technical University of Munich, Germany
(2) Department of Mathematics and Computer Science, University of Southern Denmark, Denmark
Background: Advances in omics technologies have led to massive quantitative data sets such as gene expression and networks modelling the interplay of bio-molecules such as genes, RNAs, proteins, and metabolites. Network enrichment methods combine these two data types to extract subnetworks realising systems biology responses to perturbations and diseases. However, no methods exist yet to integrate time series data with networks in a user-friendly, streamlined fashion, which severely hampers the identification of time-dependent systems biology responses.
Methods and Results: We close this gap with Time Course Network Enrichment (TiCoNE). TiCoNE combines a new human-augmented time-series clustering method with a novel approach to network enrichment. It finds temporal expression prototypes and maps them to a network of molecular interactions. The integrated data can then be investigated for enriched prototype pairs interacting more or less often than expected by chance. The significance of

these enrichments can be quantified by empirical p-values derived from fitness scores and their background distributions. Such patterns of temporal subnetwork co-enrichment can also be further compared between different conditions. With TiCoNE, we identified the first distinguishing temporal systems biology profiles in time series gene expression data of human lung cells after infection with Influenza and Rhino virus.
Discussion: TiCoNE is the first tool to allow for the identification of subnetworks enriched in biomolecules with similar time-course behaviour, the analysis of their interactions, and the comparison of identified subnetworks across different conditions. TiCoNE is available online at https://ticone.compbio.sdu.dk and as Cytoscape app at http://apps.cytoscape.org

Reducing edge noise in multi-omics correlation-based networks by fitting stochastic block models

Katharina Baum [(1)]*, Arnaud Muller [(1)], Jagath C. Rajapakse [(2)], Francisco Azuaje [(1)]
(1) Luxembourg Institute of Health, Luxembourg
(2) Nanyang Technological University, Singapore

Background: Biological entities such as genes, promoters, mRNA, metabolites or proteins do not act separately, but in concert in their network context. Edges in such molecular networks represent regulatory and physical interactions and comparing them between conditions can provide valuable information on differential mechanisms. However, biological data is inherently noisy and network reduction techniques can propagate errors particularly to the level of edges. We aim at improving the analysis of networks of biological molecules by re-establishing erroneously removed edges.
Methods and results: Our case study consists of correlation-based networks of breast cancer data stemming from high-throughput measurements of diverse molecular layers such as transcriptomics, proteomics, and metabolomics. We performed network reduction by thresholding for correlation significance or by requirements on scale-freeness. To determine missing links, we fitted the molecular networks to hierarchical stochastic block models, a method which has not yet been proposed for the analysis of correlation-based networks. For validation, we assessed the biological relevance of detected communities by functional annotations as well as the concordance of biological evidence and edge probability predictions.
Discussion: Fitting stochastic block models provides a framework to remove measurement noise in network establishment from high-throughput data. It enables deriving edge existence probabilities based on global network community characteristics, and potential hierarchies within molecular biological networks are taken into account. The edge existence probabilities could be used as an additional, integrated layer of information in network-based data comparisons.

Single cell network analysis with a mixture of Nested Effects Models

Martin Pirkl [(1)]*, Niko Beerenwinkel [(1)]
(1) ETH Zurich, Switzerland

Background: New technologies allow for the elaborate measurement of different traits of single cells under genetic perturbations. These interventional data promise to elucidate intra-cellular networks in unprecedented detail and further help to improve treatment of diseases like cancer. However, cell populations can be very heterogeneous.
Methods and results: We developed a mixture of Nested Effects Models (M&NEM) for single-cell data to simultaneously identify different cellular subpopulations and their corresponding causal networks to explain the heterogeneity in a cell population. For inference, we assign each cell to a network with a certain probability and iteratively update the optimal networks and cell probabilities in an Expectation Maximization scheme. We validate our method in the controlled setting of a simulation study and apply it to three data sets of pooled CRISPR screens generated previously by two novel experimental techniques, namely Crop-Seq and Perturb-Seq.
Discussion: M&NEMs work well in simulations and are robust even for high noise levels. Our results on the pooled CRISPR screens confirm known key regulators and we identify genes which are differently regulated among cell sub-populations.

Live analysis and privacy-preserving real-time filtering in next-generation sequencing while the sequencer is still running

Tobias P. Loka [(1)]*, Simon H. Tausch [(1)], Martin S. Lindner [(1)], Piotr W. Dabrowski [(1)], Benjamin Strauch [(1)], Jakob M. Schulze [(1)], Aleksandar Radonić [(1)], Andreas Nitsche [(1)], Bernhard Y. Renard [(1)]
(1) Robert Koch Institute, Berlin, Germany

Background: With the continuously increased use of next-generation sequencing (NGS) in time-critical applications such as disease outbreak analysis and precision medicine, there is a strong need for fast turnaround time from sample arrival to analysis results. At the same time, sequencing data in these fields may contain sensitive information that enable the re-identification of human individuals and other privacy-breaching strategies even for anonymized data. Thus, powerful methods for the analysis of NGS data that provide early interpretable results are required while also taking data protection into account.
Methods and results: We developed a collection of tools for the analysis of NGS data while the sequencer is still running. This includes software for read mapping (HiLive; Lindner et al., 2017, doi:10.1093/bioinformatics/btw659), taxonomic classification (LiveKraken; Tausch et al., 2018, doi:10.1093/bioinformatics/bty433) and pathogen identification (PathoLive). PriLive is a novel tool for the automated removal of sensitive data while the sequencing machine is running (Loka et al., 2018, doi:10.1093/bioinformatics/bty128). Thereby, reads related to a specified set of organisms of interest are unaffected. With a sensitivity of >99.4% and a specificity of >99.8% even for reads with a high number of variations, PriLive achieves results at least as accurate as conventional post-hoc filtering tools.
Discussion: With PriLive, we implemented a solution for an increased level of data protection by removing human sequence information before being completely produced. This strongly facilitates the compliance with strict data protection regulations and simplifies subsequent analyses of sensitive data.

A meta machine learning approach to distinguish true DNA variants from sequencing artefacts

Till Hartmann [(1)]*, Sven Rahmann [(2)]
(1) Genome Informatics, Institute of Human Genetics, University Hospital Essen, Germany
(2) University of Duisburg-Essen, Germany

Background: Being able to distinguish between true DNA variants and technical sequencing artefacts is a fundamental task in whole genome, exome or targeted gene analysis. Variant calling tools provide diagnostic parameters, such as strand bias or an aggregated overall quality for each called variant, to help users make an informed choice about which variants to accept or discard. Having several such quality indicators poses a problem for the users of variant callers because they need to set or adjust thresholds for each such indicator. Alternatively, machine learning methods can be used to train a classifier based on these indicators. This approach needs large sets of labeled training data, which is not easily available.
Methods and Results: The new approach presented here relies on the idea that a true DNA variant exists independently of technical features of the read in which it appears (e.g. base quality, strand, position in the read). Therefore the nucleotide separability classification problem – predicting the nucleotide state of each read in a given pileup based on technical features only – should be near impossible to solve for true variants. Nucleotide separability, i.e. achievable classification accuracy, can either be used to distinguish between true variants and technical artefacts directly, using a thresholding approach, or it can be used as a meta-feature to train a separability-based classifier. We explore both possibilities with promising results, showing accuracies around 90%.
Discussion: Using meta features (separability profiles) to train classifiers to distinguish technical artefacts from true SNVs leads to accurate models, especially given the fact that only abstract separability information is used for training.

Loss-function learning for digital tissue deconvolution

Franziska Görtler [(1)]*, Stefan Solbrig [(1)], Tilo Wettig [(1)], Peter Oefner [(1)], Rainer Spang [(1)], Michael Altenbuchinger [(1)]
(1) University of Regensburg, Germany

Gene-expression profiling of bulk tumor tissue facilitates the detection of gene regulation in tumor cells. However, differential gene expression can originate from both tumor cells and the cellular composition of the surrounding tissue. The cellular composition is not accessible in bulk sequencing but can be estimated computationally.
Digital Tissue Deconvolution (DTD) estimates the cellular composition from bulk sequencing data. Formally, DTD addresses the following problem: Given the expression profile y of a tissue, what is the cellular composition c of that tissue? If X is a matrix whose columns are reference profiles of individual cell types, the composition c can be computed by minimizing $L(y-Xc)$ for a given loss function L. Current methods use predefined all-purpose loss functions. They successfully quantify the dominating cells of a tissue, while often falling short in detecting small cell populations.
In (Görtler et al., In Proc. RECOMB 2018, 75-89, https://doi.org/10.1007/978-3-319-89929-

9_5) we use training data to learn the loss function L along with the composition c. This allows us to adapt to application-specific requirements such as focusing on small cell populations or distinguishing phenotypically similar cell populations. Our method quantifies large cell fractions as accurately as existing methods and significantly improves the detection of small cell populations and the distinction of similar cell types.

Spike-in cells enable accurate correction of contamination in single-cell transcriptomes

Nikolaus Fortelny [(1)]*, Matthias Farlik [(1)], Brenda Marquina Sanchez [(1)], Stefan Kubicek [(1)], Christoph Bock [(1)]
(1) CeMM Center for Molecular Medicine of the Austrian Academy of Sciences, Austria

Background:
Single-cell RNA-seq (scRNA-seq) is a powerful tool to study biological systems. However, biases in scRNA-seq data are poorly understood and, while computational tools are developed to align datasets and remove batch effects, little attention has been given to contamination between cells. Using spike-in cell standards with external references, we were able to assess and correct contamination in scRNA-seq data.
Methods and Results:
Mouse (32D) and human (Jurkat) cells were spiked into single-cell suspensions prior to sequencing and processed as part of each sample. Alignment to a combined mouse / human genome enabled us to accurately identify spike-in cells and to assess the extent of contamination. Strikingly, we observed a contamination of up to 20% of reads in some samples. We could further trace this contamination to cell-type marker genes that are highly expressed in the sample but absent in reference spike-in cells.
We corrected contaminated data by (i) estimating the contamination signature by comparing spike-ins to references and (ii) predicting the fraction of contaminated reads using linear models trained on spike-in cells. We further compared corrected data from our spike-in based correction method to data from a method estimating contamination without spike-ins. The spike-in based method demonstrated improved accuracy, especially in the estimation of the fraction of contaminating reads.
Discussion:
We present spike-in cells as a tool that enables accurate assessment of the extent and signature of contamination as well as computational correction of the data.

Using independent training data to improve the learning of immune cell infiltration patterns with deep learning techniques

Moritz Hess [(1)]*, Stefan Lenz [(1)], Harald Binder [(1)]
(1) Institute of Medical Biometry and Statistics Faculty of Medicine and Medical Center - University of Freiburg, Germany

Background
Tumor immune cell infiltration is a well known factor related to survival of cancer patients and is reflected in the expression of immune cell-type specific marker genes [1]. Deep learning approaches such as deep Boltzmann machines (DBMs) are capable to model complex non-

linear relations in omics data [2] and allow to generate smoothed representations of the data that may help to better uncover patterns.
Methods and results
Here we employ DBMs for modeling immune cell gene expression patterns in lung adenocarcinoma. We show that the smoothed expression data generated from a DBM allows for a better prognosis of tumor stage based on the immunome expression [3], indicating that the DBM learned a meaningful pattern which has been uncovered in [1]. Since the high number of parameters of a DBM benefits from large amounts of training data, we demonstrate how using additional expression data from other tumor entities improves the learning of meaningful patterns.
Discussion
Our results indicate that employing independent training data improves the learning of patterns in expression data, especially when the amount of training data is limited.
References
1. Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., ... & Diehn, M. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. Nature medicine, 21(8), 938-945.
2. Hess, M., Lenz, S., Blätte, T. J., Bullinger, L., & Binder, H. (2017). Partitioned learning of deep Boltzmann machines for SNP data. Bioinformatics, 33(20), 3173-3180.
3. Hess, M., Lenz, S., & Binder, H. (2018). A deep learning approach for uncovering lung cancer immunome patterns. bioRxiv, 291047.

Combining Support Vector Machines by Mixed Integer Linear Programming improves consistency of biomarker discovery in transcription profiles discriminating fungal from bacterial blood infection

Joao Saraiva [(1)]*, Rainer Koenig [(2)]
(1) Helmholtz Centre for Environmental Research - UFZ, Germany
(2) Jena University Hospital, Germany

Background: Sepsis is a life threatening disease of organ dysfunction caused by a dysregulated host response to infection. Delivery of appropriate and quick treatment to the systemic infection is mandatory and requires a quick identification of the kind of invading pathogen (e.g. fungal/bacterial). Currently, pathogen identification relies on blood cultures which take too long. The use of in situ experiments attempts to identify pathogen specific immune responses but these often lead to heterogeneous biomarkers due to the high variability in methods and materials used. Using gene expression profiles for machine learning is a developing approach to discriminate between types of infection, but also shows a high degree of inconsistency.
Methods and results : To produce consistent gene signatures, capable of discriminating fungal from bacterial infection, we have employed Support Vector Machines (SVMs) based on Mixed Integer Linear Programming (MILP). Combining classifiers by joint optimization constraining them to the same set of discriminating features increased the consistency of our biomarker list independently of leukocyte-type or experimental setup. Our gene signature showed an enrichment of genes of the lysosome pathway which was not seen by the use of independent classifiers. Moreover, our results suggest that the lysosome genes are specifically induced in monocytes. Real time qPCR of the identified lysosome-related genes

confirmed the distinct gene expression increase in monocytes during fungal infections.
Conclusion: Our combined classifier approach presented increased consistency and was able to “unmask” signaling pathways of less-present immune cells in the used datasets.

Posters

miR-QTL-Scan, a computational framework for the identification of miRNAs involved in metabolic diseases

Pascal Gottmann [(1)], Meriem Ouni [(1)], Sophie Saussenthaler [(1)], Markus Jähnert [(1)], Wenke Jonas [(1)], Matthias Blüher [(2)], Heike Vogel [(1)], Annette Schürmann [(1)]
(1) German Institute of Human Nutrition Potsdam-Rehbruecke, Germany
(2) Department of Medicine, University of Leipzig, Germany

Background: In recent studies we generated a backcross population of lean B6/C57BL/6J and obese NZO/HiDifeBom mice to determine quantitative trait loci (QTL) for obesity and diabetes. In parallel, we detected a high number of differently expressed genes (DEG) by micro array analysis of mouse strains. As most of the DEGs were located outside of QTL we hypothesized that other regulatory elements such as miRNAs are responsible for these effects.
Methods and Results: A computational framework (miR-QTL-Scan) was developed by combining QTL, genetic variants, histone modification data, miRNA target prediction tools, pathway enrichment and transcriptome analysis. Additional filtering for miRNA expression profiles identified 11 miRNAs out of 170 miRNAs in 17 QTL. Of these 11 miRNA, 9 showed a putative regulative variant. According to different target prediction approaches miR-31, located in the obesity QTL Nob6 on chromosome 4, interacts with 418 target genes in gonadal white adipose tissue (gWAT). Pathway enrichment analysis of the 418 target genes revealed a link to insulin signaling. In fact, miR-31 showed a higher expression in the gWAT of obese NZO compared to lean B6 mice, fitting to the downregulation of the insulin signaling genes. Similar patterns were detected in adipose tissue of human patients.
Discussion: In silico analysis for the detection of disease genes and in particular target prediction of miRNAs is still at an early developmental state. Nevertheless, combination of different experimental datasets, databases and pathway enrichment analysis identified suitable candidates participating in diseases development.

Construction of protein interaction network for Mycobacterium tuberculosis

Sridhar Hariharaputran [(1)]
(1) National Centre for Biological Sciences AND Indian Institute of Science AND Bharathidasan University,India, India

Background
Studying protein-protein interactions allows deeper analysis and understanding of organism's functioning, as an integrated system. A functional interaction does not need to involve direct

physical interaction or contact per se, rather it refers to a relationship between proteins that are contributing to the cellular mechanisms and through which a function is achieved by the protein. Biological function, according to Karo who argues if required then there is a need for multiple levels of context to express the biological function adequately. The intra-molecular interaction within the cell, cell-cell interactions and interactions of tissue-organ form these three different levels. The concept involved can be extrapolated to encompass the interactions between any of these sublevels and the organism. The function of the protein can be characterized by the experimentalist according to his/her perspective assuming the biochemical, cell biological and the view of genetic points. In both short term and in the long term, the notion of 'function' consists of a dynamic component, that varies over time. It is also well known that the function of protein is multifarious. Within the cell they assemble as complex and dynamic macromolecular structures, providing structural support, recognize and degrade the foreign molecules, metabolic pathways regulation, controls the progression and DNA replication through the cell cycle, synthesize other chemical species, organizing other proteins within the signal transduction cascades, mediating molecular recognition process and participate in other significant functions. And most of these processes are dealt and executed by the interacting multiprotein complexes. On the basis annotation of homologues from bacteria it is known in the pathogen Mycobacterium tuberculosis (Mtb) responsible for causing tuberculosis (TB) there are several thousand genes are known to have essential functions. To detect protein interactions there is difficulty in performing several molecular biology experiments.
Methods and results
We have used sequence, structure, ontology information along with others to generate protein-protein interaction networks and compare the interactomes based on our in-house concept/algorithm. We were able to view patterns that are followed by the participating proteins and identify hubs.
Discussion
Using computational methods to re/construct the protein interaction map of M. tuberculosis can provide the crucial information to understand the underlying biological processes in the pathogenic organism. And new therapeutic approaches can be developed based on the framework provided.

Oli2go: an automated multiplex oligonucleotide design tool

Michaela Hendling [(1)], Stephan Pabinger [(1)], Konrad Peters [(2)], Noa Wolff [(1)], Rick Conzemius [(1)], Ivan Barisic [(1)]
(1) Austrian Institute of Technology, Austria
(2) University of Vienna, Austria

Background: The success of widely used oligonucleotide-based experiments, ranging from PCR to microarray, strongly depends on an accurate design. The design process involves the selection of suitable candidates, primer pairing, specificity and primer dimer checks, applying specific parameters to produce high quality oligonucleotides. To date, many web-tools for oligonucleotide designing are available, none of which covers the whole design process.
Methods and Results: We introduce oli2go, a web-based fully automated multiplex oligonucleotide design tool. It combines all essential steps for high quality probe and primer design for a variety of biological experiments in an all-in-one solution. Especially, the oli2go

probe specificity check is not only performed against a single species (e.g. mouse), but against bacteria, viruses, fungi, invertebrates, plants, protozoa, archaea, environmental samples and sequences from whole genome shotgun sequence projects, at once. The implemented primer specificity check is performed to minimize the risk of primer binding to human background DNA and can be easily adapted to other model organisms. This allows the design of highly specific oligonucleotides in multiplex applications, which is further assured by performing dimer checks not only on the primers themselves, but in an all-against-all fashion. The software is freely accessible to all users at http://oli2go.ait.ac.at/.
Discussion: Comparing the run time of oli2go to a common oligonucleotide design workflow, we could show that oli2go is significantly faster. In contrast to a conventional oligonucleotide design process, where users must use several independent programs with limited database support or outdated algorithms, oli2go implements comprehensive algorithms and databases in one tool.

Fast Analysis of RNA-RNA Interaction Kinetics for Refining RNA Targeting Screens

Maria Waldl [(1)], Sebastian Will [(1)], Ivo Hofacker [(1)]
(1) University of Vienna, Austria

Background
The interaction of small RNAs with mRNA targets plays a major role in gene regulation. While commonly performed computational genome-wide screens identify RNA targets mostly based on thermodynamic stability and accessibility, they neglect kinetic effects, which fundamentally compromises their specificity. We identify two major obstacles to effectively integrating kinetic analysis into such screens, namely computational efficiency and modeling accuracy.
Methods and results
While single RNA folding processes have been successfully modeled as transition systems between conformations, analogous approaches for RNA-RNA interaction quickly lead to infeasibly large systems. Therefore, we propose reducing the interaction system to the direct trajectories from possible first contacts to full hybridization. Based on this key idea, we study general principles and relevant features of the interaction formation; e.g. studying the edge cases of rapid intramolecular refolding vs. much faster hybridization. Specifically, we isolate kinetic effects by comparing 62 experimentally confirmed interactions to randomized background with similar thermodynamic properties. These experiments indicate that native interactions are kinetically favored, which can be exploited to filter target predictions. Moreover, kinetics can look remarkably different dependent on the site of the initial contact.
Discussion
The proposed complexity reduction allows efficient exploration of relevant kinetic criteria. In fact, features like energy barriers can be obtained via efficient dynamic programming. This makes it possible to refine genome-wide target predictions through kinetic criteria. By studying the dependency on the interaction start site we hope to shed light on the long-debated influence of seed accessibility on interaction predictions.

Ultrafast and space-efficient k-mer indexing

Sven Rahmann [(1)]
(1) University of Duisburg-Essen, Germany

Background:
Storing sets or key-value-stores of k-mers and querying them for membership or associated values is a fundamental task for modern DNA and protein sequence analysis algorithms. This is especially true since recently, the methods that directly work on k-mers have proven to be both much faster and at least not less sensitive than classical methods based on read mapping and alignment. Applications can be found in every field; particularly well known are methods for differential gene and transcript expression analysis, metagenomics and pan-genomics (de Bruijn graphs).
Much work has focused on succinct representations of k-mer sets. However, a small memory footprint often comes at the cost of longer query times. On modern processor and cache architectures, a (random access) memory lookup can take several hundred times longer than a simple arithmetic operation. Therefore fast k-mer index data structures should minimize the number of memory accesses necessary to retrieve the queried information, ideally just a single memory lookup.
Methods & Results:
We present an engineered k-mer set and key-value store framework that achieves a good compromise between fast access and small space. Our method uses a combination of existing ideas that so far haven't been effectively combined for k-mer retrieval. It is based on three-way Cuckoo hashing onto cache-line-sized pages and stores only k-mer fingerprints large enough to ensure an exact representation. We evaluate the size of the resulting data structure for several genome-wide applications, such as xenograft sorting or a quick filter to detect reads that could indicate translocations or inversions in a human genome sample.
Discussion:
While not as space-efficient as succinct representations, the data structure we propose can still be used to represent genome-wide k-mer multisets on a typical laptop and is among the fastest of its kind.

Secure protocols for a-priori privacy checks in medical databases

Andreas Hildebrandt [(1)], Anna Katharina Hildebrandt [(2)]
(1) Institute of Computer Science, Johannes Gutenberg University Mainz, Germany
(2) MONDATA GmbH, Germany

Background:
Medical research can profit greatly from the increasing availability of large-scale data sets of medically relevant information, such as the tremendous amounts of time-resolved data generated by mobile devices and wearables. Recently, however, privacy concerns are gaining traction among users and lawmakers alike. In addition to legal requirements such as the recent European Union General Data Protection Regulation (GDPR) that must be fulfilled, highly publicized data-related scandals have made the life of database owners considerably more difficult.
Well-established privacy-preserving methods, such as differential privacy or k-anonymity, do

not seem to solve the problem. First, some aspects such as spiking in artificial noise to preserve privacy may be inappropriate in a medical context. In addition, such measures do not seem sufficient to convince users of the privacy of their data.
Methods and results:
To this end we propose a secure protocol that protects both parties – database owner and data owner – and allows a user to answer two key questions before he shares his data:
- Is his data "typical" for the data set (so that he can hide) or atypical?
- Which of the features he does not intend to disclose could be inferred from the remainder, given the data set?
We posit that these questions can be answered using modern methods from cryptography (secure multiparty computation, homomorphic encryption) without forcing the user or the database owner to disclose private data.
Discussion:
Modern protocols allow to collect large-scale medical datasets in a manner that respects privacy concerns by the data owners. We believe that in the future, methods that allow to make informed decisions about whom to share data with will become a key aspect of medical data collection.

Lyra - containerized microservices for browsing shared biomedical data

Michael Huttner [(1)], Rainer Spang [(1)], Claudio Lottaz [(1)], Stefan Hansch [(1)], Christian Kohler [(1)]
(1) Institute of functional genomics, Germany

1 Background
Research papers in the biomedical field come with large and complex data sets that are shared with the scientific community as unstructured data files via public data repositories. Examples are sequencing, microarray, and mass spec- troscopy data. Others with similar but not identical research interests can download the full data, preprocess it, integrate it with data from other publi- cations and browse those parts that they are most interested in. This requires substantial work as well as programming and analysis expertise that only few biological labs have on board.
2 Methods and results
We have developed Lyra, a collection of microservices that allows labs to make their data easily browsable over the web. Currently we provide tools for (a) insertion of genomic, proteomic, transcriptomic and metabolomic data, (b) cross linking data from different publications via automatic conversion of over 200 molecular identifier types, (c) fast data access and search over a JSON API, and (d) dynamic and interactive visualization in the users web browser.
3 Discussion
We provide software for data sharing that is highly flexible and robust, and ready for cloud computing. The split into many loosely coupled microservices allows others to take and adapt just the parts they need, and integrate them with already existing solutions. In the future we will provide full integration for kubernetes the leading container orchestration platform, making it even easier to deploy Lyra on bare-metal compute clusters or in commercial cloud solutions.

Long reads matter: The advantages of Nanopore long-read sequencing

Martin Hölzer [1]
(1) FSU Jena, Germany

Junior research group presentation

Background
In recent years, the massively parallel sequencing of DNA and cDNA (RNA-Seq) emerged as a fast, cost-effective and powerful technology to study whole genomes and transcriptomes in various ways. In the last decades, bioinformatics focused on the assessment of short-read Next-Generation Sequencing (NGS) data, most commonly derived from Illumina sequencing. Today we face a paradigm shift: long-read sequencing such as provided by Oxford Nanopore Technologies (ONT) is on the rise. By utilizing ONT, we are theoretically able to generate reads without any length cutoff. For the first time, we also have a high-throughput technology on hand that is able to sequence RNA directly - without any intermediate cDNA step.

Methods and results
The assembly of genomes and transcriptomes, still a challenging task with short-read data, will become obsolete by using ONT. We are able to distinguish and quantify different isoforms at unprecedented resolution. The technology will open up new opportunities in healthcare, clinical diagnostics and the prevention of upcoming epidemics.
Just now, the bioinformatics discipline has to balance between the comprehensive analysis of established, widely-produced, and precise short-read NGS and the novel, fast-emerging, and still more error-prone long-read ONT. Targeted algorithms and methods are needed to fully exploit the potential of this new sequencing data.

Discussion
The new junior research group headed by Dr. Martin Hölzer, which will be set up at the "RNA Bioinformatics and High-Throughput Analysis" group of Prof. Manja Marz in Jena, dares to perform this balancing act. We are going to adopt our knowledge about the comprehensive analysis of short-read NGS data and incorporate and develop new methods for long-read Nanopore assessment. Amongst others, future projects comprise 1) the characterization of differential expressed isoforms, 2) the detection of DNA/RNA modifications, and 3) the assembly-free reconstruction of viral genomes and haplotypes. Surely, massive parallel long-read sequencing will re-shape the bioinformatics way to analyze genomic and transcriptomic sequencing data for the next decades.

High genomic diversity of multi-drug resistant wastewater Escherichia coli

Norhan Mahfouz [1], Serena Caucci [2], Eric Achatz [3], Torsten Semmler [4], Sebastian Guenther [5], Thomas Berendonk [3], Michael Schroeder [3]
(1) Technical University in Dresden - TUD, Austria
(2) UNU-FLORES, Germany
(3) Technical University in Dresden - TUD, Germany
(4) Robert Koch Institute - FUB, Germany

(5) University of Greifswald, Germany

Wastewater treatment plants play an important role in the emergence of antibiotic resistance. They provide a hot spot for exchange of resistance within and between species. Here, we analyse and quantify the genomic diversity of the indicator Escherichia coli in a German wastewater treatment plant and we relate it to isolates' antibiotic resistance. Our results show a surprisingly large pan-genome, which mirrors how rich an environment a treatment plant is. We link the genomic analysis to a phenotypic resistance screen and pinpoint genomic hot spots, which correlate with a resistance phenotype. Besides well-known resistance genes, this forward genomics approach generates many novel genes, which correlated with resistance and which are partly completely unknown. A surprising overall finding of our analyses is that we do not see any difference in resistance and pan genome size between isolates taken from the inflow of the treatment plant and from the outflow. This means that while treatment plants reduce the amount of bacteria released into the environment, they do not reduce the potential for antibiotic resistance of these bacteria.

Pairs of adjacent conserved non-coding elements separated by conserved genomic distances act as cis-regulatory units

Lifei Li [(1)], Leila Taher [(1)]
(1) Friedrich-Alexander-Universität Erlangen-Nürnberg, Department of biology, Bioinformatics group, Germany

Background
Comparative genomic studies have identified thousands of conserved non-coding elements (CNEs) in the mammalian genome, many of which have been reported to exert cis-regulatory activity.
Methods and results
We analyzed ~5,500 pairs of adjacent CNEs in the human genome and found that despite divergence at the nucleotide sequence level, the inter-CNE distances of the pairs are under strong evolutionary constraint, with inter-CNE sequences featuring significantly lower transposon densities than expected. Further, we show that different degrees of conservation of the inter-CNE distance are associated with distinct cis-regulatory functions at the CNEs. Specifically, the CNEs in pairs with conserved and mildly contracted inter-CNE sequences are the most likely to represent active or poised enhancers. In contrast, CNEs in pairs with extremely contracted or expanded inter-CNE sequences are associated with no cis-regulatory activity. Furthermore, we observed that functional CNEs in a pair have very similar epigenetic profiles, hinting at a functional relationship between them.
Discussion
Taken together, our results support the existence of epistatic interactions between adjacent CNEs that are distance-sensitive and disrupted by transposon insertions and deletions, and contribute to our understanding of the selective forces acting on cis-regulatory elements, which are crucial for elucidating the molecular mechanisms underlying adaptive evolution and human genetic diseases.

Edgetic perturbation signatures reproducible across patients of a cancer type represent novel and known cancer biomarkers

Evans Kataka [(1)], Dmitrij Frishman [(1)]
(1) Techniche universität münchen, Germany

Even though researchers have made great strides in elucidating the central driver genes in cancer, determination of the complete set of cancer type (or subtype specific) and pan-cancer biomarkers is an ongoing challenge. Due to patient heterogeneity and cancer complexity as revealed by high throughput next-generation sequencing, integrating multi-omic data has been suggested as a novel way of studying the cancer microenvironment to reveal inherent biomarkers. In this experiment, we first used 642 paired (1284 in total) patient-specific non-tumour (healthy) and tumour (cancer) mRNA expression data from The Cancer Genome Atlas (TCGA) to build patient-specific protein-protein interaction networks (PPIN). Next, we compared the tumour PPIN to the corresponding non-tumour PPIN to identify molecular perturbation signatures (protein biomarkers) involved in edgetic gains or losses during tumour growth. Finally, we used 7121 known cancer-specific significantly mutated (driver) genes to determine the role somatic mutations may play in edgetic perturbations in cancer. The obtained perturbed network biomarkers were reproducible across patient samples within a cancer type (and subtype) while some biomarkers were reproducible across multiple tumours (e.g. NTRK1, MAPK4, MYOC, NUF2 and CDC45). In summary, we propose a framework to identify network biomarkers in cancer, and our results show that the approach is robust in the identification of both known and novel pan-cancer, cancer type and subtype-specific biomarkers. The identified biomarkers should be of great significance in future biomarker experimental validation, targeted therapy, and cancer monitoring.

Protein-protein interaction (PPI) network for Salmonella infected cells

Jens Rieser [(1)]
(1) Goethe-Universität Frankfurt, Germany

Salmonella typhimurium provokes gastroenteritis and typhoid fever and causes many thousands death every year. The post-translational modification of proteins after an infection is of major interest. Ubiquitination and phosphorylation are two reversible possibilities for a quick answer of the cell to environmental changes. Phosphorylation is known for activation of several protein cascades and ubiquitination builds a dense coat around cytosolic Salmonella cells [1].
To investigate the differences and similarities between ubiquitination and phosphorylation of proteins during a Salmonella infection, we analyzed the data of two different datasets, one of changes in expression levels of phosphorylated proteins in Salmonella-infected and uninfected cells [2] and the other of changes in ubiquitinated proteins [3]. We used the proteins of the two datasets to compile PPI in three databases, STRING, IntAct and BioGRID, taking into consideration the different scoring of PPI in each database.
Based on the network topology, we clustered the derived PPI network into functional groups according to their interactions applying the Girvan-Newman algorithm. To compare biological and topological clustering, we performed a GO clustering using BiNGO. To find complexes and strongly interacting subnetworks we computed cliques. Here we show different

possibilities to analyze host PPI and get better insight into the modification of proteins during a Salmonella infection. For example, HOIL1 is clustered in a group of 60 proteins, from which 27 have the GO-term 'protein ubiquitination'. In the GO clustering on the other side, it can be assigned to 'positive regulation of apoptotic signaling pathway' with a group of 31 proteins with only 35 edges, containing completey different proteins than the topological clustering. Interestingly, in both subnetworks a sixth of the proteins is only phosphorylated proteins, indicating that HOIL1 interacts mostly with ubiquitinated proteins.

A genome-wide scan for correlated mutations detects macromolecular and chromatin loop interactions in Arabidopsis thaliana.

Laura Perlaza [(1)], Dirk Walther [(1)]
(1) Max Planck Institute of Molecular Plant Physiology, Germany

Background. The concept of exploiting correlated mutations has been used in the past to efficiently identify interactions between biological macromolecules. Its rationale lies in the preservation of interactions via compensatory mutations. With the massive increase of horizontal as well as vertical sequence information, approaches based on correlated mutations have regained considerable attention.
Methods and Results. We analyzed a set of 10,707,430 single nucleotide polymorphisms detected in 1,135 accessions of the plant Arabidopsis thaliana. To measure their covariance and to reveal the global genome-wide sequence correlation structure of the Arabidopsis genome, the adjusted mutual information has been estimated for each possible pair of polymorphic sites. We developed a series of filtering steps to account for genetic linkage and/or lineage relations between Arabidopsis accessions, as well as transitive covariance as possible confounding factors. We show that upon appropriate filtering, correlated mutations prove indeed informative with regard to molecular interactions, and furthermore, appear to reflect on chromosomal loop interactions.
Discussion. Our study demonstrates that the concept of correlated mutations can also be applied successfully to within-species sequence variation and establishes a promising approach to help unravel the complex molecular interactions in A. thaliana and other species with broad available sequence information.

Using gene expression to improve precision in receptor status determination

Michael Kenn [(1)], Michael Cibena [(1)], Wolfgang Schreiner [(1)]
(1) Medical University Vienna, Austria

Background:

Receptor status is determined by immunohistochemestry to decide upon therapy in breast cancer patients. About 12% are known to be inaccurate. We present an improvement based on gene expression measurement.
Methods and results:

Gene expression data of 2880 patients are modeled by responsibility functions. These transform continous expression values into crisp receptor status, labelling certain estimates to belong to a critical domain.
Fusion of data from immunochistochemestry and gene expression modeling renders receptor status with high accuracy to be used in precision medicine.
Discussion:

We demonstrate that patients identified by our scoring might receive more adequate treatment: patients erroneously classified receptor positive but in fact negative would thus receive life saving chemotherapy. Conversely, patients considered receptor negative but truely positive might avoid chemotherapy and its side effects, since hormone therapy sufficies.

In-silico identification of genotoxic substances

Anja Friedrich [(1)], Thomas Mohr [(2)], Elisabeth Riegel [(1)], Thomas Czerny [(1)]
(1) FH Campus Wien, Austria
(2) ScienceConsult - DI Thomas Mohr KG, Austria

Genotoxic substances are chemical agents that damage the genetic information within a cell, causing mutations. Food contact materials (e.g. packaging materials) are substances intended to come directly or indirectly into contact with food. The EFSA requires the testing of food contact materials for potential genotoxic effects in two complementary tests including a bacterial gene mutation test and an in vitro micronucleus test. Recently, the European Union lowered the thresholds for genotoxic substances and as a result, most of the assays on the market are not sensitive enough, or are only able to react to a small number of substances. Human cells react with whole networks of genes to external stimuli such as genotoxic agents. Single marker genes are used to determine gene expression caused by such stimuli. Ideally such marker genes should not only react to one stimulus, but to multiple agents. The goal of this project is to find marker genes that react to as many genotoxic agents as possible in order to develop cell-based assays.
In preliminary tests multiple datasets from the GEO database have been combined into an in-house database. A Ruby pipeline was used to combine and normalise the diverse datasets to make them comparable. This led to identification of possible marker genes. In parallel, a meta analysis and GSEA were carried out to supplement the results and identify more potential target genes.
We believe that the discovery of more specific marker genes and development of cell-based assays will advance the identification of genotoxic substances present in food contact materials and medical products.

Interaction of quercetin with transcriptional regulator LasR of Pseudomonas aeruginosa: Mechanistic insights of the inhibition of virulence through quorum sensing

Hovakim Grabski [(1)], Lernik Hunanyan [(1)], Susanna Tiratsuyan [(1)], Hrachik Vardapetyan [(1)]
(1) Russian-Armenian University, Armenia

Pseudomonas aeruginosa is one of the most dangerous superbugs for which new antibiotics are urgently needed. This bacterium forms biofilms that increase resistance to antibiotics and host immune responses. Current therapies are not effective because of biofilms. Biofilm formation is regulated through a system called quorum sensing, which includes transcriptional regulators LasR and RhIR. These transcriptional regulators detect their own natural autoinductors. Thus disrupting this system is considered a promising strategy to combat bacterial pathogenicity. It is known that quercetin inhibits Pseudomonas aeruginosa biofilm formation, but the mechanism of action is unknown. In the present study, we tried to analyse the mode of interactions of LasR with quercetin.
We used a combination of molecular docking, molecular dynamics simulations and machine learning techniques, which includes principal component and cluster analysis, to study the interaction of the LasR protein with quercetin. We show that quercetin has two binding modes. Both binding modes are not competitive. One binding mode is the interaction with ligand binding domain, which has been shown experimentally. The second binding mode is the interaction with the bridge. It involves conserved amino acid interactions from multiple domains. This part has not been shown experimentally, because LasR protein is not soluble. But biochemical studies show hydroxyl group of ring A is necessary for inhibitory activity. In our model the hydroxyl group interacts with multiple leucines during the second binding mode. This study may offer insights on how quercetin inhibits quorum sensing circuitry by interacting with transcriptional regulator LasR.

Machine Learning and Mycobacterium tuberculosis

Sridhar Hariharaputran [(1)]
(1) National Centre for Biological Sciences AND Indian Institute of Science AND Bharathidasan University,India, India

Background
Machine-learning approach can take advantage of the data generated from WGS to develop new antibiotics combining with genomics. Further the cost and time to generate the data is also decreasing. Apart from the FDA approved drug targets against Mycobacterium tuberculosis there is a shortage of compounds that are yet to be directed towards new targets. Previously computational methods such as homology modeling and docking have been used to identify active molecules.
Methods and results
We try to apply machine learning methods to bridge the gaps or for connecting the dots with Mycobacterium tuberculosis and other Mycobacterium species and identify the patterns and compare and analyse their genomes. Earlier we have integrated large data sets and built separate data resources such as MyGOnets, MyCompare, MyPockets, SinCRe each one of them focusing on different aspects of Mtb and interactions. Using this variety of information and the pattern we try to spot the unknown and their features. We are able to identify some and assign or connect the features using our methods.
Discussion
An integrated method can help in construction of interactomes, build consensus and compare features and drug target identification.

Tracing the genomic footprint of natural competence in bacterial genomes

Sachli Zafari [(1)], Holger Bergmann [(1)], Ingo Ebersberger [(2)]
(1) Dep. for Applied Bioinformatics Goethe University Frankfurt, Germany
(2) Dep. for Applied Bioinformatics Goethe University, Senckenberg Biodiversity and Climate Research Centre, Frankfurt, Germany

Background
Horizontal gene transfer (HGT) is a potent evolutionary process for bacteria to accomplish genetic innovation. Among all HGT mechanisms, natural competence (NC), the capability to uptake and integrate free DNA, is the most versatile way to uptake novel genes. Yet, only a small number of bacteria appears constitutively naturally competent, but the number of bacteria with a hitherto unobserved conditional NC might be substantially larger. Here, we investigate if NC shows an observable footprint along a genome that could be exploited to assess the prevalence of NC in bacteria.
Methods and Results
We first determine whether or not a species is capable of uptaking free DNA, in principle. To this end, we trace known competence machineries across the bacterial domain using a domain-architecture-aware targeted ortholog search. We then apply HGTector followed by a validation of HGT candidates via phylogenetic profiling to quantify number and genomic distribution of horizontally acquired genes in species displaying a competence machinery, and in such that don't. Our results reveal an overall high number of horizontally acquired genes in constitutively competent bacteria, whereas non-competent species display fewer HGTs with a tendency for local clustering.
Discussion
Recent studies link human pathogenicity and antibiotic resistance to NC. But to our knowledge, no approach can differentiate competent from non-competent organisms without experimental validation. We show that the acquisition of numerous genes is more prevalent and dispersed in competent than in non-competent bacteria which can serve as a signal to evaluate in-silico the capability of uptaking free DNA.

Composing a dockerized Ecosystem for the Exchange and Visualization of Biological Networks

Florian Auer [(1)], Frank Kramer [(1)], Tim Beissbarth [(2)]
(1) University Medical Center Göttingen, Germany
(2) University Medicine Göttingen, Germany

Integration of biological networks into data analyses are common techniques within bioinformatics workflows, but still face problems in reproducibility, particularly in situations that build upon collaborative work. Within the process, the visualization of networks has huge impact on the interpretation, and therefore is an essential step in communicating even intermediate results. Furthermore, the shift of bioinformatics towards systems medicine, and its application within a clinical setting demand the establishment of interconnected and interchangeable components and, moreover, a standardized interface for information

exchange with healthcare systems.

We demonstrate a course from data acquisition to the visualization, while providing an interface for healthcare systems by exposing Fast Healthcare Interoperability Resources (FHIR®). Thereby the network data exchange (NDEx) platform and the Cytoscape project form the core components.
We use our R package ndexr to retrieve networks from public and private NDEx installations, and also to store the results for collaboration and publication. Cytoscape is a major tool for the visual exploration of biomedical networks. Beside the graphical interface, it can be accessed by the R package RCy3. To reduce the complexity of software installations, we utilize docker to compose the different components, illustrating their interchangeability and flexibility furthermore.

Bioinformatics analysis identifies the renoprotective factor dicarbonyl and L-xylulose reductase (DCXR) as prognostic chronic kidney disease biomarker

Paul Perco [(1)], Wenjun Ju [(2)], Julia Kerschbaum [(3)], Johannes Leierer [(3)], Matthias Kretzler [(2)], Gert Mayer [(3)], Michael Rudnicki [(3)]
(1) Medical University of Innsbruck, Austria
(2) University of Michigan, United States
(3) Medical University Innsbruck, Austria

Background: An imbalance of nephroprotective factors and renal damaging molecules contributes to development and progression of chronic kidney diseases. In this study we determined the potential of renoprotective factors to predict disease progression in a set of chronic kidney disease (CKD) patients.
Methods and Results: Gene expression profiles were determined for 197 previously published renoprotective factors in 63 CKD patients. The statistical analysis of microarray (SAM) method was used to identify downregulated factors in the group of progressive patients. Progression was defined as reaching end-stage renal disease or doubling of serum creatinine. Significance of renoprotective factors to predict course of disease was in addition analysed in time-to-event analysis using Kaplan Meier curves and log-rank statistics. Cox regression models were used to adjust for the clinical parameters estimated glomerular filtration rate (eGFR) and diagnosis type.
The six renoprotective factors DCXR, EGF, GSTM1, KNG1, NOS3, and UMOD were significantly associated with disease outcome. DCXR (p-val = 0.0277), EGF (p-val = 0.0264), GSTM1 (p-val = 0.0039), and KNG1 (p-val = 0.0372) remained significant after adjustment for eGFR and diagnosis in multivariate Cox regression analysis. Downregulation of DCXR on the transcriptional level was validated in two independent CKD cohorts. This is to our knowledge the first study describing the prognostic potential for DCXR in human CKD samples with literature evidence already being available for the other three markers showing significance in our cohort.
Conclusion: We showed that downregulation of the renoprotective factor dicarbonyl and L-xylulose reductase (DCXR) is an independent predictor of disease outcome in chronic kidney disease.

Complexity and Function in the Human Genome

Anton Pirogov [(1)], Peter Pfaffehluber [(2)], Angelika Börsch-Haubold [(3)], Bernhard Haubold [(3)]
(1) Max-Planck-Institute for Evolutionary Biology, Plön, Germany
(2) Freiburg University, Germany
(3) Max-Planck-Institute for Evolutionary Biology, Germany

Background
Approximately half the human genome consists of transposon remnants. We therefore suggest that the genome reflects a transposon mutagenesis experiment carried out over evolutionary time. Under this scenario, regions sensitive to structural change should be free of recent transposon insertions. The aim of this project is to rapidly identify such regions and to test whether they are enriched for certain genetic functions.
Methods & Results
We assess the absence of recent transposon insertions by measuring sequence complexity. This is high in the absence of sequence duplication and low otherwise. Our program macle implements a measure of complexity based on match lengths observed in a sliding window. Matches can occur anywhere in the genome. Macle first computes a permanent index of the human genome in 3.5 hours, which is then traversed in 25 seconds.
We find that regions with complexity indistinguishable from random are twofold enriched for promoters. More significantly, we find that these highly complex regions are up to 40-fold enriched for developmental genes.
Discussion
High complexity is equivalent to low similarity to the rest of the genome. Our premise is to regard the human genome as the result of a transposon mutagenesis experiment, where recent transposon insertions leave footprints of low complexity. The finding that complex regions are enriched for developmental genes demonstrates the fruitfulness of this heuristic. We next aim to extend this complexity/function analysis to fully sequenced genomes from across the tree of life.

Superiority of multiple random gene sets in predicting survival of patients with hepatocellular carcinoma

Timo Itzel [(1)], Rainer Spang [(2)], Thorsten Maass [(3)], Stefan Munker [(4)], Hans Hürgen Schlitt [(5)], Wolfgang Herr [(6)], Matthias Evert [(7)], Andreas Teufel [(1)]
(1) Division of Hepatology, Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Germany
(2) Statistical Bioinformatics, Department of Functional Genomics, University Medical Center, Regensburg, Germany
(3) Hepacult GmbH, Regensburg, Germany

(4) Department of Medicine II, Großhadern University Medical Center, Ludwig Maximilians University, Germany
(5) Department of Surgery, University Medical Center, Regensburg, Germany
(6) Department of Medicine III, University Medical Center, Regensburg, Germany
(7) Department of Pathology, University of Regensburg, Germany

Despite multiple publications, molecular signatures predicting the course of hepatocellular carcinoma (HCC) have not yet been integrated into clinical routine decision making. Given the diversity of signatures published throughout the past decade, optimal number, best combinations, and benefit of functional associations of genes in prognostic signatures still remain to be defined.
We investigated a vast number of randomly chosen gene sets (varying between 1 and 10 000 genes) in order to encompass the full range of prognostic gene sets on 242 transcriptomic profiles of patients with HCC.
Depending on the selected size, 4.7 to 23.5% of all random gene sets exhibit prognostic potential by separating patient subgroups with significantly diverse survival. This was further substantiated by investigating gene sets and signaling pathways also resulting in a comparable high number of significantly prognostic gene sets.
However, combining multiple random gene sets using “swarm intelligence” resulted in a significantly improved predictability for approximately 63% of all patients. In these patients, approx. 70% of all random 50-gene containing gene sets resulted in equal and stable prediction of survival. For all other patients a reliable prediction seems highly unlikely for any selected gene set. Using a machine learning and independent validation approach, we demonstrated a high reliability of random gene sets and swarm intelligence in HCC prognosis.
In conclusion, we demonstrate that using “swarm intelligence” of multiple gene sets for prognosis prediction may not only be superior but also more robust for predictive purposes.

Hepamine - A Liver Disease Microarray Database, Visualization Platform and Data-Mining Resource

Timo Itzel [(1)], Matthias Evert [(2)], Andreas Teufel [(1)]
(1) Division of Hepatology, Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Germany
(2) Department of Pathology, University of Regensburg, Germany

Motivation: Throughout the past two decades, numerous gene expression profiling data on literally all liver diseases were generated and stored in public databases. These data are thought to contain deep insights into the molecular development of liver diseases, support the development of molecular diagnostics and ultimately promote precision medicine in hepatology. However, once published the majority of these data remain idle. Only very few data were used for additional analyses or comparative projects by the hepatology research community. This may mostly be due to the limited bioinformatics knowledge on how to obtain and analyze the stored raw data by most biomedical research personnel. In order to overcome this barrier and to support an easy translation of bioinformatics data into translational hepatology research, we created Hepamine, a liver disease microarray database, visualization platform and data-mining resource.
Methods: Microarray data were obtained from the ArrayExpress Archive of Functional

Genomics Data (http://www.ebi.ac.uk/arrayexpress). Pre-analysis of expression data was performed using R statistical software and microarray analysis packages from the Bioconductor repository (https://www.bioconductor.org). Expression data were stored locally in a postgreSQL database.
Results: We have generated Hepamine, a web-based repository of pre-analyzed microarray data for various liver diseases. Among these are HCC, CCC, liver fibrosis/cirrhosis, chronic hepatitis, autoimmune liver disease, fatty liver disease and many more. At its initial release Hepamine contains 14 gene expression datasets, 21 microarray experiments and roughly half a million gene expression measurements. A self-explanatory website offers open and easy access to the respective gene expression profiles and support to compare samples, experiments and various liver diseases. Genes may be searched on the basis of specific expression patterns across diverse samples. Results are furthermore visualized in simple three color tables indicating up-, down-, or no differential expression in multiple experiments. To enlarge the scope of the Hepamine, all data were linked to common functional and genetic databases, in particular offering information on the respective gene, signaling pathway analysis and evaluation of biological functions by means of gene ontologies.
Conculsion: Hepamine provides comprehensive data and easy access to various hepatologic gene expression data. It will open this widely unused resource particularly to hepatologists without bioinformatics or microarray profiling experience and substantially facilitate the translation of these data to molecular hepatology research. Hepamine is accessible at: http://www.hepamine.de.

Measurable Residual Disease monitoring with next-generation sequencing

Razif Gabdoulline [(1)], Felicitas Thol [(1)], Alessandro Liebich [(1)], Piroska Klement [(1)], Johannes Schiller [(1)], Martin Wichmann [(1)], Blerina Neziri [(1)], Konstantin Büttner [(1)], Bennet Heida [(1)], Arnold Ganser [(1)], Michael Heuser [(1)]
(1) Hannover Medical School, Germany

One third of patients with acute myeloid leukemia (AML) develops a clinical relapse after allogeneic hematopoietic stem cell transplantation (alloHSCT). Measurable residual disease (MRD) monitoring may detect molecular signs of relapse prior to clinical relapse in patients without morphologic/microscopic evidence of disease. This was clearly shown, for example, by MRD assessment of mutated nucleophosmin 1. However, the majority of AML patients do not have this marker, therefore we developed an MRD workflow using alternative MRD markers that can be measured by next-generation sequencing (NGS).
We designed over 150 primers targeting short (100-150 bp) genomic regions in order to quantify the genetic variations found in an initial mutation screening by a custom True Sight Myeloid Sequencing Panel.The primer design makes use of UMI-s (universal molecular identifiers - short random sequences attached at the early stages of sequencing), which label predominantly the sequences originating from the same DNA molecule. This approach significantly increases the variant detection sensitivity of MRD monitoring. Bioinformatics challenges are: quality control of sequencing results, efficient UMI detection and clustering, development of robust and sample quality-dependent methods for quantifying the limit of detection.
The achieved detectable variant allele fraction was ~ 0.02% compared to ~1% in sequencing without error correction. The clinical significance of the approach was evaluated in 116 AML

patients undergoing alloHCT in complete morphologic remission. Suitable mutations for MRD assessment were found for 93% of the patients. 5-year incidence of relapse for MRD-positive patients was significantly higher than for MRD-negative patients, 66% vs 17%.
NGS-based MRD monitoring requires a sophisticated experimental design and bioinformatics analysis, but in return it is widely applicable to AML patients, highly predictive of relapse and survival, and may help refining transplant and post-transplant management in AML patients.

Independent component analysis provides insights into biological processes and clinical outcomes for melanoma patients

Petr Nazarov [(1)], Anke Wienecke-Baldacchino [(2)], Andrei Zinovyev [(3)], Urszula Czerwinska [(3)], Arnaud Muller [(1)], Gunnar Dittmar [(1)], Francisco Azuaje [(1)], Stephanie Kreis [(2)]
(1) Luxembourg Institute of Health, Luxembourg
(2) University of Luxembourg, Luxembourg
(3) Institut Curie, France

Background
The amount of publicly available cancer-related "omics" data is constantly growing and can be used to gain insights into cancer biology of the new patients and their diagnosis. However, the integration of different data sets is not straightforward and requires special approaches to deal with heterogeneity at technical and biological levels. Here we develop a methodology that can mitigate technical biases and improve our understanding of the cancer biology in the new coming patients.
Methods and results
The approach is based on the independent component analysis (ICA) – a data-driven method of signal deconvolution. We implemented parallel consensus ICA that robustly decomposes transcriptomic data into independent signals or components. This separates the true signals of distinct cell types from technical biases. The weights of independent components were used for prediction of clinically relevant patient characteristics, such as tumour subtype or patient survival. Moreover, the components were linked to biological functions and the new samples were characterized by the presence of the respective biological properties such as immune response, angiogenic activity or cancer cell proliferation. Finally, ICA-based integration of transcriptome and miRNome data allowed deducing biological functions of miRNAs, which would not be possible otherwise.
Discussion
Taken together, ICA represents a versatile tool to dissect complex data into individual components. It allows for better integration of public and new-coming data, thus valorising large cancer datasets obtained over the past years. The components can be linked to biological processes, cell types and clinically relevant outputs: tumour type and patient survival. In case when several layers of "omics" data are available, ICA can be used to link subsets of features such as genes, miRNAs and proteins to a single biological process.

Towards a Systematic Understanding of Drug-Drug Interactions

Michael Caldera [(1)], Felix Müller [(1)], Jörg Menche [(1)]
(1) CeMM - Research Center for Molecular Medicine of the Austrian Academy of Sciences, Austria

Drug-drug interactions (DDIs), i.e. changes in the effect of a drug when used in combination with another drug, have important implications for clinical applications (e.g. the efficacy of a treatment, side effects) as well as for drug development (e.g. combination therapies). Although the concept of DDIs has been known for nearly a century, relatively little is known about general rules and patterns underlying such drug combination effects on the cellular (cell autonomous) level. To fill this gap, we have analysed a high-throughput high-content imaging screen of 35.778 drug combinations from a representative library of FDA-approved drugs. The images were created and analysed using state of the art fluorescent imaging techniques as well as image analysis technology, including machine learning for feature selection and image quality control. Applying a novel methodology based on high-dimensional cell morphology feature vectors extracted from these images, allowed us to identify the full extent of reciprocal and joint interactions between drugs and ultimately to create a comprehensive DDI network of unprecedented resolution, including fully weighted and directed interactions. Conducting an analysis of the resulting DDI network alone, as well as an integrative analysis in the context of other molecular and phenotypic networks, such as protein-protein interaction networks, pathway maps and disease networks, helped us understanding how and when interactions occur, by revealing their molecular mechanisms. Overall, this project represents a novel framework for the extraction and application of high dimensional cell morphology feature vectors from fluorescent imaging data in the context of DDIs and is a first systematic attempt to reveal the fundamental arithmetics of drugs, i.e. a profound, molecularly rooted and predictive understanding of how the effects of individual drugs add up when used in combination.

Bioinformatic Analysis of FXR related ChIP-seq data

Emilian Jungwirth [(1)], Katrin Panzitt [(2)], Martin Wagner [(2)], Gerhard G. Thallinger [(1)]
(1) Institue of Computational Biotechnology, Graz University of Technology, Austria
(2) Divison of Gastroenterology and Hepatology, Medical University Graz, Austria

Background: Chromatin immunoprecipitation sequencing (ChIP-seq) is a method to identify genome-wide transcription factor (TF) binding sites and to gain information about transcriptional regulation. The TF Farnesoid X receptor (FXR) is a nuclear receptor (NR) that controls gene regulation of different metabolic pathways in the liver (e.g. bile acid, lipid and glucose metabolism) and recently attracted a lot of attention as novel drug target for various metabolic liver diseases.
Aims: Several FXR ChIP-seq data sets for different species, conditions and cell lines have been reported, but none so far for human liver tissue. Our aim is to re-analyze publicly available FXR ChIP--seq data sets with one standardized method and compare them to our own human FXR-liver tissue data sets.
Methods: We searched in public-sources for available FXR-ChIP-seq data sets to determine a common set of generally accepted quality criteria based on the ones proposed in the

ENCODE guidelines and other authoritative ChIP-seq guidelines. Furthermore we investigated different parameter settings and input variants. Ultimately, based on common quality and analysis criteria a comparison between the different data sets will be made including further downstream analysis.
Results: In public resources there are three FXR-ChIP-seq data sets available for mice, one for rat and one for human primary hepatocytes. Most data sets include baseline FXR binding and binding events under pharmacological treatment or diseased conditions. No data sets are available for human liver tissue. Our ongoing analysis shows that the data sets we have re-analyzed so far are heterogeneous in regard to quality criteria, which also includes our own data set of human liver biopsy samples. Additionally, we observe that the analysis results are sensitive to settings of certain parameters and the choice of input.
Conclusion: Several FXR-ChIP-seq data sets are publicly available for various species and conditions. Standard Encode quality criteria are usually not reported for those data sets. A meta-analysis of these different data sets with standardized methods should help to get a comprehensive and global overview of FXR binding events and FXR-dependent gene regulation across various species.

Stratification of macrolide treatment for community-acquired pneumonia

Xueqi Cao [(1)], Christina Forstner [(2)], Marcus Oswald [(1)], Mathias Pletz [(2)], Rainer König [(1)]
(1) System Biology, Center of Sepsis Control and Care, Jena University Hospital, Jena, Germany
(2) Institute of Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

BACKGROUND:
Community-acquired pneumonia (CAP) remains a disease with considerable morbidity and mortality (9.5% day-180 mortality) [1]. Antibiotic therapy administering macrolides are very effective in many cases, but rather ineffective in other cases or even detrimental, as they may lead to cardiotoxic side-effects in susceptible patients [3]. Early treatment decision is mandatory for a successful clinical case and it’s a challenge to decide the appropriate therapy immediately at consultation (day-0).
In this study, we aim to find etiological and clinical parameters serving as theranostic markers to support macrolide treatment decision.
METHOD AND RESULT:
The data from the CAPNETZ study were analyzed [1,2]. As this was not a randomized controlled clinical trial for macrolides treatment, by employing Generalized Linear Models, a propensity score was calculated for each patient to balance health conditions between treated /non-treated patients [4]. Using simple Fisher's Exact Tests followed by decision trees, the criteria for the subgroup, which benefit most from macrolides treatment, were identified.
DISCUSSION:
The most promising features favoring macrolide treatment included no chronical co-morbidities and specific breath rate range. Together, they formed a subgroup, in which the death rate was considerably reduced from 9.0% to 2.4% if they were treated according to our rule. Furthermore, with a more complex rule obtained from the decision trees, the death rate was reduced by 2.7% in patients who were in compliance with the rule, compared to patients who were not.

Overall, macrolides could be efficient antibiotics to treat CAP if the appropriate group of patients is selected.

Dynamics of MinHash Values on Genomic q-Gram Sets

Henning Timm [(1)], Sven Rahmann [(1)]
(1) University of Duisburg-Essen, Germany

Background
MinHashing is used to compute the similarity of sets using their minimal hash values, e.g. between the q-grams of a read and those of a reference sequence. It relies on the assumption that the hash values generated from elements of the sets are chosen without bias. However, biological sequences can deviate greatly from a uniform distribution, which can influence the efficiency of MinHashing approaches.

Methods and Results
The hash repeat length (HRL), i.e. the number of q-grams after which the minimal hash value changes, is an important metric to judge how well a sequence is suited for MinHashing analysis. The HRL can be restricted to be smaller than a window size w to enable comparison of sequences with different sizes. We analyzed the distribution of HRLs for reference genomes, as well as simulated sequences with different GC content, to judge the impact of the parameters q and w. Following, we formulated a model to predict the distribution of hash repeat values.

Discussion
Knowing the HRL distribution allows to compensate for genome specific deviations when using MinHashing. We also aim to provide suggestions for best practices to use MinHashing in combination with other techniques. For example, canonical q-grams (the minimum of a q-gram and its reverse complement under a numerical representation) are often used to mitigate the influence of reverse complementary sequences. While this technique is widely used, the interplay of MinHashing and canonical q-grams is not well analyzed yet, raising the question: Can using maximal canonical q-grams improve the performance of MinHashing approaches?

In-depth analysis of immunological and genetic tumor status in and across cancer types: impact of mutational signatures beyond total mutational load

Jan Budczies [(1)], Anja Seidel [(2)], Eugen Rempel [(1)], Petros Christopoulos [(1)], Barbara Seliger [(3)], Peter Schirmacher [(1)], Albrecht Stenzinger [(1)], Carsten Denkert [(2)]
(1) Universitätsklinikum Heidelberg, Germany
(2) Charité - Universitätsmedizin Berlin, Germany
(3) Martin-Luther-University Halle-Wittenberg, Germany

Background
Harnessing the immune system by checkpoint blockade has greatly expanded the therapeutic

options for advanced cancer. Since the efficacy of immunotherapies is influenced by the molecular make-up of the tumor and its crosstalk with the immune system, comprehensive analysis of genetic and immunologic tumors characteristics is essential to gain insight into mechanisms of therapy response and resistance.
Methods and results
We investigated the association of immune cell contexture and tumor genetics including tumor mutational burden (TMB), CNA load, mutant allele heterogeneity (MATH) and specific mutational signatures (MutSigs) using TCGA data of 5722 tumor samples from 21 cancer types. Among all genetic variables, MutSigs associated with DNA repair deficiency and AID/APOBEC gene activity showed the strongest positive correlations with immune parameters. For smoking-related and UV-light-exposure associated MutSigs a few positive correlations were identified, while MutSig 1 (age related/clock-like process) correlated non-significantly or negatively with the major immune parameters in most cancer types. High TMB was associated with high immune cell infiltrates in some but not all cancer types, in contrast, high CNA load and high MATH were mostly associated with low immune cell infiltrates. While a bi- or multimodal distribution of TMB was observed in some cancer types including colorectal and stomach cancer where it was associated with MSI status, a non-dichotomized analysis of TMB appeared more appropriate for many other cancer types including NSCLC and melanoma.
Discussion
This study uncovered specific genetic-immunology associations in major cancer types and identified mutational signatures as interesting candidates for response prediction beyond TMB.

The evolutionary traceability of proteins

Arpit Jain [(1)], Fabian Fliedner [(1)], Arndt von Haeseler [(2)], Ingo Ebersberger [(1)]
(1) Goethe University Frankfurt, Germany
(2) Centre for Integrative Bioinformatics Vienna, Austria

Background
Orthologs document the evolution of genes and metabolic capacities encoded in extant genomes. Orthologous genes detected in all domains of life allow reconstructing the gene set of LUCA, the last universal common ancestor. These genes inform about the functional repertoire common to – and necessary for – all living organisms. Recently, a minimal gene (MG) set for a self-replicating cell was determined experimentally. Surprisingly many genes have unknown functions and are not represented in LUCA. However, as similarity between orthologs decays with time, it becomes insufficient to infer common ancestry. Thus, ancient gene set reconstructions are incomplete and distorted to an unknown extent.
Methods and Results
Here we introduce the evolutionary traceability, together with the software protTrace, that quantifies, for each protein, the evolutionary distance beyond which the sensitivity of the ortholog search becomes limiting. protTrace estimates for a seed protein its specific evolutionary rate together with constraints on the evolutionary change jointly from a pre-compiled ortholog set and from the seed protein’s domain architecture. A simulation based framework estimates then the traceability decay with time. We show that the LUCA set comprises only highly traceable proteins. Furthermore, proteins in the MG set lacking

orthologs outside bacteria mostly have low traceability. Thus, their eukaryotic orthologs might have been overlooked. We demonstrate how a traceability-informed adjustment of the search sensitivity identifies hitherto missed orthologs.
Discussion
The evolutionary traceability helps to differentiate between true absence and non-detection of orthologs, and thus improves our understanding about the evolutionary conservation of functional protein networks.

Deep neural network approach to predict clinical outcomes in high-dimensional and smallsample data settings

Leon-Charles Tranchevent [(1)], Francisco Azuaje [(1)], Jagath Rajapakse [(2)]
(1) Luxembourg Institute of Health (LIH), Luxembourg
(2) Nanyang Technological University, Singapore

Background
The availability of high-throughput omics datasets for large patient cohorts has allowed the development of methods that aim at predicting patient clinical outcomes such as survival and disease recurrence. Such methods are also important to better understand biological mechanisms underlying disease etiology and development. In this context, we investigated a deep learning strategy for clinical outcome prediction. One of the main challenge of such strategy is the «small n large p» problem. Omics datasets typically consist of few samples and numerous features relative to typical deep learning datasets. Neural networks usually tackle this problem through feature selection or by including additional constraints during the learning process.
Methods
We propose a novel strategy that relies on a graph-based method for feature extraction, coupled with a deep neural network for clinical outcome prediction. The expression data are first represented as graphs whose nodes represent patients, and edges represent correlations between the patients' expression profiles. Topological features, such as centralities, are then extracted from these graphs for every node. Lastly, these features are used as input to train and test neural networks. We explore how different parameters and layers of the network are selected in order to overcome the effects of small data problem as well as the curse of dimensionality.
Results
We apply this strategy to four neuroblastoma datasets and observe that models based on neural networks are more accurate than state of the art models (DNN: 85%-87%, SVM/RF: 70%-81%). We also tested our models using independent datasets, generated using different gene expression platforms. Our results indicate that the artificial neural networks capture complex features in the data that help predicting patient clinical outcomes.

Structural and functional factors explain the observed difference in sequence conservation between transmembrane and soluble domains.

Mark Teese [(1)], Peter Hönigschmid [(1)], Martin Ortner [(1)], Dominik Müller [(1)], Shenger Wang [(1)],

Rima Jeske [(1)], Dmitrij Frishman [(1)], Dieter Langosch [(1)]
(1) Technical University of Munich, Germany

Background: In membrane proteins, residue conservation is known to be higher for transmembrane (TM) domains than soluble extramembrane (EM) domains. However, the underlying structural and functional factors are poorly understood.
Methods and results: TM/EM conservation ratios were calculated for non-redundant datasets of α-helical membrane proteins from humans, yeast, and available crystal structures. Homologues were obtained by BLAST, strictly filtered, and TM/EM conservation ratios calculated using an identity matrix. Our results show that single-pass proteins have much lower relative TM conservation than multi-pass proteins. We show that enzymes have low TM/EM conservation ratios. In fact, the purifying selection of catalytic domains is so strong that for single-pass enzymes, TM regions are much less conserved than EM regions. We also discuss the evolutionary patterns of other functional groups, including ion channels, GPCRs, immune proteins, and transporters. Using available crystal structures, we show that normalisation for residue burial and “random conservation” removes all differences between TM and EM regions.
Discussion: We propose that exposure of the polar backbone of polypeptides is unfavourable within the apolar lipid environment, causing TM regions to be highly compact. The higher propensity for residue burial within these compact domains explains the majority of difference between TM and EM regions. Single-pass proteins and enzymes have low TM/EM conservation ratios because the propensity for residue burial is high within their large soluble domains, and low within their TM domains. After accounting for residue burial, TM regions still appear more conserved due to the overabundance of hydrophobic Leu, Ile, Phe, and Val residues in alignments. This concept of “random conservation” is relevant to any protein region with a limited amino acid propensity.

PCRedux: machine learning helper tool for sigmoid curves

Michal Burdukiewicz [(1)], Andrej-Nikolaj Spiess [(2)], Stefan Rödiger [(3)]
(1) University of Wrocław, Poland
(2) Department of Andrology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, Germany
(3) BTU Cottbus - Senftenberg, Germany

Background: Quantitative Real-Time PCR (qPCR) has a considerable popularity as a simple and robust method in research and precision medicine. It is an important tool to counterconfirm NGS experiments. Software packages and workflows for processing (e.g., normalization procedures, relative gene expression analysis) qPCR data exit. However, accurate unbiased and reproducible classification (e.g., negative, ambiguous, positive) of samples from high-throughput runs requires an automated qPCR curve identification method capable of labeling unknowns against a large cohort of known samples. There is no open source software package that contains classified data sets and provides biostatistical methods for machine learning on amplification curves.
Methods and results: Herein, we describe the `PCRedux` package (https://github.com/devSJR/PCRedux), which is an add-on package (MIT license) for open source statistical computing language and environment R. `PCRedux` contains methods for

automated qPCR curve classification based on feature calculation and machine learning methods. A large collection of amplifications curves (n = 50000) was compiled. We used `PCRedux` for a population wide microRNA screening. Our technology was tested in silico through multiple cross-validations using amplification curves from different experimental conditions. We demonstrated over 95% accuracy. In vitro verification was done using qPCR runs. A web server prototype is available from http://www.smorfland.uni.wroc.pl/shiny/predPCR/. The web server uses the machine learning `mlr` package as interface to a large number of classification and regression techniques. Discussion: The machine learning approach enabled reliable, scalable, and automated qPCR curve classification with broad potential clinical and epidemiological applications.

Big Data and Complex Network Analysis in Bionformatics

Blagoj Ristevski [(1)]
(1) Faculty of Information and Communication Technologies – Bitola, University “St. Kliment Ohridski” – Bitola, Macedonia

Abstract This research aims to obtain novel knowledge from modeling and analysis of biological complex networks based on available large amount of biological data.
BACKGROUND: Nowadays, big data analysis in biomedicine is very promising process of integrating, exploring and analysing of large amount complex heterogeneous data with different nature: biomedical data, experimental data, EHRs and social media data. These types of data provide very suitable basis to construct biological complex networks. A complex network is a graph consisted of nodes and edges, with non-trivial characteristics. Dependending on the edges in the graph structure, networks can be directed or undirected, weighted or unweighted. To understand bilogical characteristics of the complex networks quantitatively and qualitatively, network topology properties are crucial.
METHODS AND RESULTS: In this research, several types of the complex networks are covered such as: gene regulatory networks, gene co-expression networks, metabolic networks and protein-protein interactions networks.The properties of these biological complex networks, such as node degrees, in- and out - degree distribution of the nodes, assortativity, betweenness, centrality, clustering coefficient, connectedness, path length of the networks, eccentricity, diameter of a graph, cliques and networks motifs, modularity, network topology and robustness are analyzed.
DISCUSSION: The obtained knowledge is beneficial to reveal the new properties of the biological complex networks, network motifs and hubs, as well as examination of the networks robustness and their behavior.

Entropic hourglass patterns of animal and plant development

Alexander Gabel [(1)], Hajk-Georg Drost [(2)], Marcel Quint [(1)], Ivo Grosse [(1)]
(1) Martin Luther University Halle-Wittenberg, Germany
(2) University of Cambridge, United Kingdom

One surprising observation going back to pioneering works of Karl Ernst von Baer in 1828

and Ernst Haeckel in 1866 is that embryos of different animal species express on average evolutionarily young genes at the beginning of embryogenesis, evolutionarily old genes in mid-embryogenesis, and again evolutionarily young genes at the end of embryogenesis. Focusing our attention on plants, which represent the second major kingdom in the tree of life that evolved embryogenesis, we have found that this phylotranscriptomic hourglass pattern also exists in plant embryogenesis, which is surprising as multicellularity and embryogenesis evolved independently in animals and plants. Moreover, we have found that phylotranscriptomic hourglass patterns also exist in the two main transitions of post-embryonic plant development, germination and floral transition, suggesting the convergent evolution of phylotranscriptomic hourglass patterns in animal and plant development.
The origin of these hourglass patterns has remained concealed, but here we find that not only the mean age of expressed genes changes in an hourglass-like manner, but the whole age distribution of expressed genes changes. When studying the entropy of these age distributions as functions of time, we find hourglass patterns that surprisingly are orders of magnitude more significant than the original hourglass patterns of the mean, which might indicate that the entropic hourglass patterns are more fundamental than, and possibly even the origin of, the original hourglass patterns of animal and plant development.

A k-mer based Deep Learning Approach for Quantitative, Reference-free Prediction of Antibiotic Resistance Phenotypes from Next-Generation Sequencing Data

Thomas Walsh [(1)], Marko Fritz [(1)], Stephan Beisken [(1)], Andreas E Posch [(1)]
(1) Ares Genetics GmbH, Austria

Background

Antimicrobial resistance (AMR) is a global health threat with a projected annual casualty rate of more than 10 Mio by 2050. Monitoring the spread of AMR via in vitro testing is time consuming and costly. In silico predictions of quantitative AMR phenotypes, such as the minimum inhibitory concentration (MIC), can therefore greatly facilitate AMR monitoring. Current reference-based, qualitative methods depend on closely related reference genomes for genetic AMR marker identification and predict AMR from published clinical breakpoints for the categorization of sample MICs into resistant or susceptible. Reference-free, quantitative workflows have the potential to detect unknown causal genetic AMR markers missed by reference-based approaches and the prediction of MICs – rather than resistance labels – enables more precise AMR monitoring.

Methods and Results

We use k-mers – short nucleotide or protein sequences of length k – that enable reference-free detection of any genomic variation in a set of samples to build an accurate MIC prediction pipeline with deep learning. Deep learning is able to build high-resolution abstractions for increased predictive performance and differentiate relevant features without upfront feature engineering.
Both deep learning and k-mer analysis have been used to address several aspects of the problem of antimicrobial resistance, but to date they have not been used in combination to predict MIC values. We trained the models on the proprietary ARESdb comprising a representative set of thousands of thoroughly profiled clinical isolates from different pathogens that have been collected globally over the last 30 years. Our approach allows reliable MIC prediction across multiple pathogens within the limits of natural biological

variation across MIC tests.

Essential Bioinformatics Web Services for Sequence Analyses

Robert Penchovsky [(1)], Nikolet Pavlova [(1)]
(1) Sofia University, Bulgaria

We are going to present new Essential Bioinformatics Web Services (EBWS) are implemented on a new PHP-based server that provides useful tools for analyses of DNA, RNA, and protein sequences applying a user-friendly interface. 13 Web-based applets are currently available on the Web server. They include reverse complementary DNA and random DNA/RNA/peptide oligomer generators, a pattern sequence searcher, a DNA restriction cutter, a prokaryotic ORF finder, a random DNA/RNA mutation generator. It also includes calculators of melting temperature (TM) of DNA/DNA, RNA/RNA, and DNA/RNA hybrids, a guide RNA (gRNA) generator for the CRISPR/Cas9 system and an annealing temperature calculator for multiplex PCR and several other useful bioinformatics applets (http://penchovsky.atwebpages.com/applications.php).
Part of this research is published in IEEE Transactions on Computational Biology and Bioinformatics(EBWS: Essential Bioinformatics Web Services for Sequence Analyses. IEEE Transactions on Computational Biology and Bioinformatics DOI: 10.1109/TCBB.2018.2816645). The research in Penchovsky laboratory is supported by a grant, number DN13/14/20.12.2017 awarded by the Bulgarian National Science Fund.

Constraining maximal intermolecular helix lengths improves RNA-RNA interaction prediction

Rick Gelhausen [(1)], Sebastian Will [(2)], Ivo L. Hofacker [(2)], Rolf Backofen [(1)], Martin Raden [(1)]
(1) University of Freiburg, Germany
(2) University of Vienna, Austria

Background
The computational prediction of RNA-RNA interactions based on thermodynamic models is widely used to identify targets of regulatory RNAs. While tools typically can predict the interaction of RNAs accurately, they predict many highly-ranked false positives in screens for RNA interaction targets. An important source of this low specificity is that the current secondary-structure-based models do not reflect many steric constraints that govern the kinetic formation of RNA-RNA interactions. For example, often short kissing hairpin interactions cannot be easily extended, since this would require sterically prohibited unwinding of intramolecular helices. In consequence, methods that do not consider such effects predict over-long helices.
Methods and results
To increase the prediction accuracy, we suggest to length-restrict the runs of consecutive intermolecular base pairs (perfect stackings) in the successful state-of-the-art IntaRNA algorithm. By preventing the prediction of long helices, this considers steric and kinetic aspects, which could not be efficiently modeled otherwise. We devise the efficient dynamic programming recursions and extend our IntaRNA implementation. The helix-lengths-

constrained approach is thoroughly benchmarked on a large prokaryotic data set for sRNA target prediction. We show that the restriction of intermolecular helix lengths significantly improves the prediction accuracy and lowers IntaRNA's runtime.
Discussion
The sterically/kinetically motivated restriction of intermolecular helix lengths can be easily included in existing RNA-RNA interaction prediction tools and enhances their prediction accuracy. While we introduced the method for IntaRNA, i.e. for an accessibility-based RNA-RNA interaction prediction model, the idea and results are generic and can be transferred to other models as well.

Learning the Topology of Latent Signaling Networks from High Dimensional Transcriptional Intervention Effects

Zahra Sadat Hajseyed Nasrollah [(1)], Achim Tresch [(1)], Holger Fröhlich [(2)]
(1) Institute of Medical Statistics and Computational Biology(IMSB), Germany
(2) Bonn-Aachen International Center for Information Technology (B_IT), Germany

Data based learning of the topology of molecular networks, e.g. via Dynamic Bayesian Networks (DBNs) has a long tradition in Bioinformatics. The majority of methods take gene expression as a proxy for protein expression in that context, which is principally problematic. Further, most methods rely on observational data, which complicates the aim of causal network reconstruction. Nested Effects Models (NEMs – Markowetz et al., 2005) have been proposed to overcome some of these issues by distinguishing between a latent (i.e. unobservable) signaling network structure and observable transcriptional downstream effects to model targeted interventions of the network.
The goal of this project is to develop a more principled and flexible approach for learning the topology of a dynamical system that is only observable through transcriptional responses to combinatorial perturbations applied to the system. More specifically, we focus on the situation in which the latent dynamical system (i.e. signaling network) can be described as a network of binary state variables with logistic activation functions. We show how candidate networks can be scored efficiently in this case and how topology learning can be performed via Markov Chain Monte Carlo (MCMC).
As a first step, we extensively tested our approach by applying it to several known network motifs (Feed Forward, Feed Backward, Bifan, Diamond, Protein Cascading) over a wide range of possible settings (e.g. different number of observations, time points). Moreover, we evaluated our method with synthetic data generated from ODE systems taken from the literature. As a next step we evaluate our method with data from the DREAM 8 challenge.
In future work, we also plan to extend our method to incorporate multi-omics data and apply it to patient samples to identify disease related networks.

EBWS: Essential Bioinformatics Web Services for Sequence Analyses

Nikolet Pavlova [(1)], Dimitrios Kaloudas [(2)], Robert Penchovsky [(1)]
(1) Sofia University, Bulgaria
(2) Mediterranean Agronomic Institute of Chania, Greece

In this poster, we present Essential Bioinformatics Web Services (EBWS) implemented on a new PHP-based server that provide user-friendly interface and useful tools for analyses of DNA, RNA, and protein sequences. Nine new Web-based applets are already available on the Web server and published (Kaloudas, D. and colleagues, EBSW: Essential Bioinformatics Web Services for Sequence Analyses, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 15, 2008). EBWS include reverse complementary DNA and random DNA/RNA/peptide oligomer generators, a pattern sequence searcher, a prokaryotic ORF finder, a DNA restriction cutter, a pattern sequence searcher, and a random DNA/RNA mutation generator. The pattern-searching applet has no limitations in the number of motif inputs. It applies a toolbox of Regex quantifiers used for defining complex sequence queries of RNA, DNA, and protein sequences. The DNA enzyme digestion program utilizes a large database of 1,502 restriction enzymes. To them, the server includes calculators of melting temperature of DNA/DNA, RNA/RNA, and DNA/RNA hybrids, an annealing temperature calculator for multiplex PCR, and a guide RNA (gRNA) generator for the 9 CRISPR/Cas9 system. The gRNA generator has a database of 25 bacterial genomes and it is searchable for gRNA target sequences. It has an option for searching in any genome sequence given by the user. There are 4 additional still unpublished applets such as DNA/RNA translator, protein sequence code translator from three letters to one letter and vice versa, Virtual PCR and hydropathy plots of proteins. All thirteen programs are freely available online at http://penchovsky.atwebpages.com/applications.php.
This research is supported by grant DN13/14/20.11.2917 awarded by the Bulgarian National Science Fund and by grant project BG05M2OP001-2.009-0019-C01/02.06.2017, financed by the Operational Program ‘Education and Science for Intelligent Growth’, co-financed by the European Union through the European Structural and Investment Funds.

Predicting potential drug-drug interactions using statistical learning

Andrej Kastrin [(1)], Polonca Ferk [(1)], Brane Leskošek [(1)]
(1) Institute of Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia

Background
Drug-drug interaction (DDI) is a change in the effect of a drug when patient takes another drug. Characterizing DDIs is extremely important to avoid potential adverse drug reactions. We represent DDIs as a complex network in which nodes refer to drugs and links refer to their potential interactions. Recently, the problem of link prediction has attracted much consideration in scientific community.
Methods and results
We represent the process of link prediction as a binary classification task on networks of potential DDIs. We use link prediction techniques for predicting unknown interactions between drugs in five arbitrary chosen large-scale DDI databases, namely DrugBank, KEGG, NDF-RT, SemMedDB, and Twosides. We estimated the performance of link prediction using a series of experiments on DDI networks. We performed link prediction using unsupervised and supervised approach including classification tree, k-nearest neighbors, support vector machine, random forest, and gradient boosting machine classifiers based on topological and semantic similarity features. Supervised approach clearly outperforms unsupervised

approach. The Twosides network gained the best prediction performance regarding the area under the precision-recall curve (0.93 for both random forests and gradient boosting machine).
Discussion
The applied methodology can be used as a tool to help researchers to identify potential DDIs. The supervised link prediction approach proved to be promising for potential DDIs prediction and may facilitate the identification of potential DDIs in clinical research.

The Swiss Molecular Tumor Board: Comprehensive Molecular Cancer Diagnostics in the Clinics

Franziska Singer [(1)], Anja Irmisch [(2)], Nora Toussaint [(3)], Linda Grob [(4)], Jochen Singer [(5)], Thomas Thurnherr (5), Niko Beerenwinkel [(5)], Mitch Levesque [(6)], Reinhard Dummer [(6)], Luca Quagliata [(7)], Sacha Rothschild [(8)], Andreas Wicki [(9)], Christian Beisel [(10)], Daniel Stekhoven [(5)]
(1) ETH Zurich; NEXUS Personalized Health Technologies, Switzerland
(2) University of Zurich Hospital, Switzerland
(3) Clinical Bioinformatics Unit, NEXUS Personalized Health Technologies, ETH Zurich, Switzerland
(4) ETH, Switzerland
(5) ETH Zurich, Switzerland
(6) University of Zurich, Switzerland
(7) Pathology, University Hospital Basel, Switzerland
(8) Oncology, University Hospital Basel, Switzerland
(9) usb, Switzerland
(10) D-BSSE, ETHZ, Switzerland

Molecular profiling of tumors based on high-throughput techniques has become an emerging practice in hospitals all over the world. Transitioning from direct testing of a few specific targets to the analysis of comprehensive high-throughput data provides numerous benefits to patients. This is particularly true for the treatment of patients with rare diseases, patients with tumors lacking known targetable mutations, patients with cancer of unknown primary and for patients with limited treatment options.
Despite this great potential, the application of comprehensive molecular diagnostics is not yet well established in clinical practice, largely due to the challenge of accurate and robust bioinformatics analysis and clinically meaningful variant interpretation. Further, such pipelines require specific protocols that account for stringent quality control, privacy issues, and thorough process documentation.
To this end, we combined the bioinformatics expertise of the Swiss Federal Institute of Technology (ETH) platform NEXUS Personalized Health Technologies with the clinical expertise of the University Hospitals in Zurich and Basel to implement a workflow for molecular diagnostics of cancer patients. Matched tumor and normal samples undergo comprehensive high-throughput sequencing and are analyzed with a focus on the identification of actionable variants to improve clinical decision support. We combine whole-exome, whole-genome, and transcriptome sequencing to investigate somatic changes on the single nucleotide, copy number, and expression level. The detected changes are linked to possible treatments and clinical trial options. The results are summarized in a concise and clearly structured clinical report designed to facilitate discussions in a clinical molecular tumor

board.

The m6A reader protein YTHDC2 interacts with the small ribosomal subunit and the 5′-3′ exoribonuclease XRN1

Jens Kretschmer [(1)], Harita Rao [(2)], Philipp Hackert [(1)], Katherine Sloan [(1)], Claudia Höbartner [(2)], Markus Bohnsack [(1)]
(1) University Medical Centre Göttingen, Germany
(2) Georg-August-University, Germany

N6-methyladenosine (m6A) modifications in RNAs play important roles in regulating many different aspects of gene expression. While m6As can have direct effects on the structure, maturation or translation of mRNAs, such modifications can also influence the fate of RNAs via proteins termed “readers” that specifically recognise and bind modified nucleotides. Several YTH domain-containing proteins have been identified as m6A readers that regulate the splicing, translation or stability of specific mRNAs. In contrast to the other YTH domain-containing proteins, YTHDC2 has several defined domains and here, we have analysed the contribution of these domains to the RNA and protein interactions of YTHDC2. The YTH domain of YTHDC2 preferentially binds m6A-containing RNAs via a conserved hydrophobic pocket, whereas the ankyrin repeats mediate an RNA-independent interaction with the 5′-3′ exoribonuclease XRN1. We show that the YTH and R3H domains contribute to the binding of YTHDC2 to cellular RNAs and using crosslinking and analysis of cDNA (CRAC), we reveal that YTHDC2 interacts with the small ribosomal subunit in close proximity to the mRNA entry/exit sites. YTHDC2 was recently found to promote a “fast-track” expression programme for specific mRNAs and our data suggest that YTHDC2 accomplishes this by recruitment of the RNA degradation machinery to regulate the stability of m6A-containing mRNAs and by utilising its distinct RNA-binding domains to bridge interactions between m6A-containing mRNAs and the ribosomes to facilitate their efficient translation.

Kernelized Rank Learning for Personalized Drug Recommendation

Lukas Folkman [(1)], Xiao He [(2)], Karsten Borgwardt [(2)]
(1) CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Austria
(2) ETH Zurich, Switzerland

Background: Large-scale screenings of cancer cell lines with detailed molecular profiles against libraries of pharmacological compounds are currently being performed in order to gain a better understanding of the genetic component of drug response and to enhance our ability to recommend therapies given a patient's molecular profile. These comprehensive screens differ from the clinical setting in which (1) medical records only contain the response of a patient to very few drugs, (2) drugs are recommended by doctors based on their expert judgment, and (3) selecting the most promising therapy is often more important than accurately predicting the sensitivity to all potential drugs. Current regression models for drug sensitivity prediction fail to account for these three properties.

Methods and results: We present a machine learning approach, named Kernelized Rank Learning (KRL), that ranks drugs based on their predicted effect per cell line (patient), circumventing the difficult problem of precisely predicting the sensitivity to the given drug. Our approach outperforms several state-of-the-art predictors in drug recommendation, particularly if the training dataset is sparse, and generalizes to patient data.
Discussion: Our work phrases personalized drug recommendation as a new type of machine learning problem with translational potential to the clinic.

Flexible and interactive visualization of GFA sequence graphs

Giorgio Gonnella [(1)], Niklas Niehus [(1)], Stefan Kurtz [(1)]
(1) University of Hamburg, Center for Bioinformatics (ZBH), Germany

Background
GFA (https://github.com/GFA-spec) is an emerging format for representing sequence graphs, including assembly de Bruijn and string graphs and variant graphs. It is supported by sequence assemblers, read mappers, sequence variant analysis tools, and scripting language libraries, such as GfaPy (Gonnella and Kurtz, 2017) and RGFA (Gonnella and Kurtz, 2016). Bandage (Wick et al., 2015), an interactive visualization tool for assembly graphs, supports, among other formats, GFA1. However, the latest version of the format specification (GFA2) introduced new powerful features which allow extending the use case of the format to e.g. mapping and assembly of long reads, scaffolding graphs and representation of variant graphs.
Methods and Results
Here, we present GfaViz, a tool for the interactive visualization of GFA sequence graphs. The tool was implemented in C++ based on the Qt framework and the OGDF library. It allows the user to select among different force-based layout algorithms. The visual representation of the graph can be fully customized (e.g. colors, proportions, font of single elements or the whole graph) and exported to vector and raster image formats.
Discussion
GfaViz is, to our knowledge, the first visualization tool supporting the GFA2 format, including all new features of the revised format, such as the support of generalized local alignments (i.e. not necessarily end-to-end), representation of alignments of reads to contigs, gaps, and subgraphs including any specified subset of the graph. The layout and customizations are saved in the GFA file itself as application-specific metainformation. Thus no external configuration files are required.

Transcriptome and proteome dynamics during pollen development and their visualization through VoronoiTreemaps

Stefan Simm [(1)], Mario Keller [(2)], Ken Oehler [(2)], Henning Wehrmann [(2)], Enrico Schleiff [(3)]
(1) Goethe University and FIAS, Germany
(2) Goethe University, Germany
(3) Goethe University, FIAS and BMLS, Germany

Background – Pollen development is the central step of plant reproduction and has been analysed extensively by means of transcriptomic or proteomic approaches. However, so far a combined multi-omics approach is still missing. To address this issue, we combined a NGS-based transcriptomics and a mass spectrometry-based proteomics approach to examine the transcriptome and proteome of tomato pollen at tetrad, post-meiotic and mature stage.
Methods and Results – The analysis of the transcriptome and proteome size during the development of pollen revealed an increase in transcriptome complexity from early to late stages with a concomitant decrease in proteome complexity. Further, a quantitative analysis led to the identification of two translation modes active in pollen, which were termed direct and delayed translation. For the functional analysis of the transcriptomes and proteomes we developed VoronoiTreeraiser, including a user-friendly GUI to create and adjust Voronoi Treemaps based on GO, KEGG or KOG functional hierarchies and expression data from -omics experiments. The reported Voronoi Treemaps represent the expression data in a functional context and by this provide direct information about active processes.
Discussion – Our results provide novel insights in transcriptome and proteome dynamics during pollen development. These dynamics include an uncoupling of the transcriptome and proteome, which was on the one hand apparent in the different behaviour of transcriptome and proteome complexity during pollen development and on the other hand by the observed delayed translation of certain transcripts. Further, with VoronoiTreeraiser we present a universally applicable tool for a combined quantitative and functional analysis.

Efficient calculation of microbial production envelopes

Sarah Noel Galleguillos [(1)], Matthias Gerstl [(1)], Jürgen Zanghellini [(1)]
(1) Austrian Centre of Industrial Biotechnology, Austria

Phenotypic phase planes, sometimes called production envelopes, are an important tool in constraint-based analysis of metabolic networks. They allow one to characterize the full metabolic capabilities of an organism as function of selected reaction fluxes of interest. However, phase planes are most often evaluated for only two fluxes of interest (typically a product of interest as function of growth), as the computational work load scales exponentially with the number of selected fluxes.
Here we present an efficient algorithm for the fast calculation of phase planes in multi dimensions in any genome-scale metabolic model. We use concepts developed in computational geometry to efficiently enumerate the vertices of the phase plane, resulting from the projection of the total metabolic capabilities onto the dimensions of interest (biomass, product and substrate fluxes). With our approach, phase plane analysis is no longer restricted to two or three reactions of interest. This is particularly important for the unbiased analysis of microbial communities. In fact, for the first time, it is now becoming possible to use phase planes to characterize multiple interactions and dependencies between members of a microbial community.

Correspondence analysis of human brain organoid transcriptome data

Ela Gralinska [(1)], Naresh Mutukula [(1)], Sneha Arora [(1)], Yechiel Elkabetz [(1)], Martin Vingron [(1)]
(1) Max Planck Institute for molecular Genetics, Germany

Cerebral organoids – miniature brain organs created in vitro – can be a valuable tool for studying brain development and various neurological disorders. However, most of existing methods for deriving such systems from pluripotent stem cells (PSCs) are highly variable and lead to heterogeneous neural cell populations. Hence, a gold standard protocol for deriving homogeneous populations is needed.
To tackle this problem, here we compared organoids derived by existing protocols to human brain tissues and assessed to which extent each of the resulting organoids resemble the different brain regions tested. We integrated RNA-seq data of organoids derived using different protocols with RNA-seq data of the developing human brain from the Allen Brain Atlas. Cortical identity of organoid samples was investigated using correspondence analysis (CA). CA, similarly to principal component analysis (PCA), is a method for projecting a data matrix into a low dimensional subspace. In contrast to PCA, CA can simultaneously account for the samples in the gene-dimensional space and genes in the sample-dimensional space, and highlight associations between genes and samples. We identified a significant experimental bias between organoid and human brain datasets, which we removed by applying the ComBat function from the sva package. The resulting normalized data was again subjected to CA, yielding a plot adjusted for the observed batch effect. This plot could be interpreted in terms of the cortical identity of the investigated samples.
On a computational level, our approach gives an example of how to integrate and visualize transcriptome data from two different sources, while biologically our analysis provides a rational basis for comparison of protocols for deriving homogenous cortical cell populations.

ChimeraMATE: An algorithm for the de-novo detection of chimeric DNA sequences

Marius Welzel [(1)], Manfred Jensen [(2)], Jens Boenigk [(2)], Sven Rahmann [(3)], Daniela Beisser [(1)]
(1) University of Duisburg-Essen, Genome Informatics / Biodiversity, Germany
(2) University of Duisburg-Essen, Biodiversity, Germany
(3) University of Duisburg-Essen, Genome Informatics, Germany

A chimeric sequence is a nucleic acid fragment composed of sub-fragments that originated from two or more parent sequences. In environmental amplicon studies, where closely related but phylogenetically distinct sequences are amplified together, the formation of chimeric sequences that do not occur under natural conditions in an organism impose a major bioinformatical challenge, as they can lead to heavy distortions of the detected operational taxonomic units (OTUs).
In this study we developed a new algorithm for the detection of chimeras in sequencing datasets that does not rely on chimera-free reference databases. Instead, we opted for a de-novo graph-based approach, deconstructing the sequences into individual k-mers to find subsequences shared by two or more sequences. By solving the shortest common superstring problem and using sequence abundances we can draw directed acyclic graphs representing the relationship between sequences to deduce if a sequence is chimeric or not. To account for naturally occurring ubiquitous k-mers, e.g. highly conserved regions, soft-

masking is used to exclude these regions from the analysis.
Initial tests show promising results, with a detection rate of 70-84 % and a false positive rate of 5-8 % in simulated chimera data sets.

New Approaches for Meta-Analysis and Network Meta-Analysis of Transcriptomics Data

Robin Kosch [(1)], Christine Winter [(1)], Klaus Jung [(1)]
(1) University of Veterinary Medicine Hannover, Germany

Background
High-dimensional transcriptome expression data from public repositories can be utilized in meta-analysis and network meta-analysis to increase statistical power, to gain further knowledge with a higher level of scientific evidence and to make indirect inferences between study groups. Two new concepts are presented, here.
Methods and results
Direct ('early') data merging followed by a joint analysis of selected gene expression studies can be an alternative to meta-analysis by ('late') merging of results. While several methods for meta-analysis of differential expression studies have been proposed, meta-analysis of gene set enrichment tests have very rarely been considered. In this work, we compare different strategies of meta-analysis of gene set tests. In simulation studies and in an example of manipulated real world data, we found that in most scenarios the early merging has a higher sensitivity of detecting a gene set enrichment than the late merging.
Network Meta-analysis aims to make indirect group comparisons that have not been considered in the original studies. So far, network meta-analysis has not been considered to make indirect comparisons in transcriptome expression data, when data merging appears to yield biased results. In simple study networks the results of the network meta-analysis highly correlates with the results from merged data.
Discussion
Research synthesis by meta-analysis can make result of transcriptomics data analysis more robust. However, depending on a concrete study scenario, analysts carefully have to decide between data and results merging.

Differential snoRNA expression in miRNA sequencing experiments

Stephan Bernhart [(1)], Stephanie Kehr [(1)]
(1) Leipzig University, Germany

Background: SnoRNAs are a class of noncoding RNAs that guide post-transcriptional RNA modifications. Differential usage of snoRNAs has been described in various cases such as cancer and development. Accordingly, finding differentially expressed snoRNAs is important. The median length of a common snoRNAs is 79nt for box C/D and 133nt for box H/ACA snoRNAs, respectively. In total the sequences range from 24 to 420nt. Meaning that snoRNAs are typically but not necessarily longer than microRNAs. However, as there is an abundance of miRNA sequencing experiments, we ventured to find out whether these experiments can also be used to investigate differential snoRNA expression.

Method: We used a dataset generated in the development of iPSC cells to hepatocyte like cells where 50nt of a small RNA fraction were sequenced. Most sequenced reads did not contain adapter sequences, showing that the original sequences were longer than 50nt. We compared the differential usage of short snoRNA reads with the differential usage of the total snoRNA reads and found that as a whole, there is very good correlation between these signals.
Discussion: Thus miRNASeq experiments can be used to analyze differential snoRNA usage.

The RNA workbench

Joerg Fallmann [(1)], Florian Eggenhofer [(2)]
(1) University of Leipzig, Germany
(2) University of Freiburg, Germany

RNA centric research is of growing importance for medicine and molecular biology. Increasing amounts of data from deep sequencing experiments create a demand for automatic analysis and interpretation solutions.
The RNA-Workbench offers a wide range of tools covering classic RNA-bioinformatics as well as RNA-seq fields. Predefined workflows for the annotation of non-coding RNAs or identification of differentially expressed genes are subsets of over 50 included tools from the categories RNA structure analysis, RNA alignment, RNA annotation, RNA-protein interaction, ribosome profiling, RNA-seq analysis and RNA target prediction. RNA specific visualisation solutions for dot-bracket plots and secondary structures are part of the workbench.
In contrast to pre-existing solutions, our community driven approach allowes us to include classic RNA-bioinformatics tools often with the direct support of the tool-authors. These contributions enable us to provide excellent documentation, training material and interactive tours demonstrating the functionality of the workbench.
Building on the Galaxy framework the workbench offers sophisticated analyses to users without command line knowledge, while emphasising reproducibility, customization and effortless scale up to larger infrastructures. The workbench is implemented as Galaxy Docker flavour and therefore easily extendable by additional tools, workflows, tours or training data, that can be installed from the Galaxy ToolShed. The workbench will be further improved and maintained in an ongoing community effort.

Origin of Enhancer Redundancy

Nicolai Barth [(1)], Leila Taher (1)
(1) FAU Erlangen-Nürnberg, Germany

Background
Many gene regulatory networks appear to contain enhancers with partially overlapping

activity. Those groups of enhancers are referred to as redundant or shadow enhancers. It is silently assumed, that shadow enhancers mainly originate by means of duplication. It is, however, also possible that two shadow enhancers originate independently, for instance via independent transposon insertion and subsequent refunctionalization events. Here, we investigated whether independent origin of shadow enhancers may be more widespread than generally assumed.
Methods and results
We utilized the set of human enhancers predicted by the FANTOM5 project. First, we assigned target genes to the enhancers via correlation of activity patterns of enhancers and promoters. Then, we grouped the enhancers according to their target genes and activity patterns to identify groups of redundant enhancers. Redundant enhancers show features that differ significantly from non-redundant enhancers. For example, redundant enhancers are less conserved. Approximately half of the enhancers overlap with transposon annotation, pointing at the importance of transposon amplification in enhancer evolution. Moreover, when compared in a pairwise manner, the enhancers in a group of redundant enhancers overlap with different types of transposons, and thus, appear to have originated independently form one another.
Discussion
Our results provide evidence that independent origin is indeed a widespread mechanism of shadow enhancer evolution.

Multi-omics and time-resolved data integration for drug response analysis

Arnaud Muller [(1)], Evelyn Ramberger [(2)], Katharina Baum [(1)], Petr Nazarov [(1)], Francisco Azuaje [(1)], Gunnar Dittmar [(1)]
(1) Luxembourg Institute of Health, Luxembourg
(2) Max Delbrück Center for Molecular Medicine, Germany

Background
The emergence of different high-throughput omics technologies led to a tremendous increase of data in regard to scale and diversity. Integrating these diverse data poses a critical challenge for data processing and concurrently an opportunity for biological discovery.
Methods and results
By creating features representing multi-omics profiles – transcriptomics and proteomics – in two different conditions (drug treatments) in time resolution, we propose here a method that combines the tSNE (t-Distributed Stochastic Neighbor Embedding) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithms to visualize and cluster these features.
In this study we analyze acute promyelocytic leukemia cell lines (NB4) treated with either 12-O-tetradecanoyl-phorbol-13-acetate (TPA) or with all-trans-retinoic-acid (ATRA) leading to differentiation into monocytes or granulocytes respectively. Transcriptome (microarray) and proteome (SILAC) data was collected as a time series (7 time-points for ATRA, 5 for TPA and the starting condition -T0). Following significant differential expression analysis -using T0 as the common reference- genes and proteins expressions were merged into features by matching gene symbols (Entrez) and protein names (UniProt). Thus, each feature summarizes the combined expression profile for the transcriptome and proteome for both drug treatment across time-points.

This data integration generated a list of more than 5000 features, which were analyzed by tSNE followed by HDBSCAN allowing the visualization of (dis)similar behaviors during differentiation. Biological relevance of clusters was assessed by functional analysis using Gene Ontology and Reactome databases.
Discussion
Feature engineering followed by machine learning techniques allowed relevant insights in high-level multi-dimensional heterogeneous data by taking into account, as a whole, the most biologically meaningful data. Further investigation will help to elucidate the complex systems of cell differentiation.

Integration of multi-omics lung cancer data through a patient-oriented and pathway-centric approach

Sang Yoon Kim [(1)], Arnaud Muller [(1)], Tony Kaoma [(1)], Petr Nazarov [(1)], Victoria El Khoury [(1)], François Bernardin [(1)], Gunnar Dittmar [(1)], Francisco Azuaje [(1)]
[(1)] Luxembourg Institute of Health (LIH), Luxembourg

BACKGROUND: In the personalized medicine era, the accurate identification and classification of cancer patients is a major challenge. In addition, various robust omics methods have been introduced which allow scientists to interpret molecular mechanisms of cancers using multiple types of “omics” dataset. Therefore, the integration of multi-omics data is an indispensable step to develop new approaches to the clinically-relevant stratification of patients and support the investigation of new therapies.
METHOD AND RESULTS: Here we identified patient subgroups of lung cancer by integrating multi-omics datasets using SNF (Similarity Network Fusion) and single sample GSE (Gene-Set enrichment) methods. We retrieved lung cancer patient data from a Luxembourg lung cancer cohort (n=45), which contains protein expression (mass spectrometry and tissue microarray experiments) and DNA mutation profiles (Ion Ampliseq Cancer Hotspot Panel). We also analyzed RNA-Seq, protein expression (RPPA) and methylation profiles from The Cancer Genome Atlas (TCGA LUAD, n=318). First, we constructed patient similarity networks with such omic datasets, and then combined these networks using SNF method. In addition, using GSE methods, we mapped single sample-level data onto gene sets, and performed the same network analysis. Next, we applied spectral clustering to the fused networks, and investigated the association between the clusters and patient survival using Cox log-rank test. In the TCGA data, we found that the GSE-based analyses led to significantly differential survival between patient subgroups (p-value<0.01).
DISCUSSION: GSE can provide us with stronger and clearer interpretations about clinically-relevant molecular profiles by incorporating biological knowledge into the analysis. In comparison to gene-centric analyses, GSE-based approach allows the identification of clusters with stronger differential survival. However, gene-centric methods can still provide useful information. Therefore, both approaches could play complementary roles in personalized medicine.

DEUS: an R package for accurate small RNA profiling based on differential expression of unique sequences

Tim Jeske [1], Peter Huypens [2], Laura Stirm [3], Selina Höckele [2], Christine Wurmser [4], Anja Böhm [5], Cora Weigert [3], Harald Staiger [3], Johannes Beckers [2], Maximilian Hastreiter [6]
(1) Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München GmbH, Germany
(2) Institute of Experimental Genetics, Helmholtz Zentrum München GmbH, Germany
(3) Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Zentrum München at the University of Tübingen, Germany
(4) Chair of Animal Breeding, Technische Universität München, Wissenschaftszentrum Weihenstephan, Germany
(5) Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Zentrum München at the University of Tübingen, Germany
(6) Institute of Computational Biology, Helmholtz Zentrum München GmbH, Germany

Since their discovery, a growing number of small non-coding RNA (sncRNA) classes has emerged. Despite their fundamental role in various biological processes, the analysis of sncRNA sequencing data remains a challenging task. Mapping of sequenced fragments to a reference genome is a widely used practice to compile fragments originating from the same RNA molecule and to identify their genomic location. Here, we show that mapping-based approaches are forced to cause data loss or uncontrolled shifts in expression intensities and present our R-package DEUS providing a novel strategy for accurate sncRNA profiling.
We have collected a set of 150 sncRNA sequencing samples from human and murine biomaterials. We demonstrate that multi-mapping is a considerable problem affecting approximately 60% of all input reads. We observed that multi-mapping is significantly more abundant in biomaterials with a potential carrier function when compared to somatic cell types and colorectal cancer cell lines. In order to avoid mapping-related problems, we applied DEUS to these data sets. Our pipeline first creates a unique sequence based count table and then performs differential expression analysis. Finally, sequences are annotated via BLAST and clustered according to their sequence similarity.
DEUS analyzes sncRNA sequencing data without requiring a reference genome and without causing any data loss. Still, it enables an intuitive feature-based interpretation of differential expressed sncRNAs. Additionally, results encompass qualitative differences between experimental conditions which would likely be masked by mapping-based approaches. DEUS is implemented as a highly modular open-source R package and freely available at http://ibis.helmholtz-muenchen.de/deus/.

Differential analysis of combinatorial protein complexes with CompleXChange

Thorsten Will [1], Volkhard Helms [1]
(1) Universität des Saarlandes, Germany

Background:
Many proteins operate in multiprotein complexes and not on their own. Unlocking this complexome in a condition-specific manner thus promises a deeper understanding into the cellular wriring and what happens upon cell fate transitions. Although there exist large

amounts of transcriptomic data and an increasing amount of data on proteome abundance, quantitative knowledge on the dynamics of complexomes is lacking.
Methods and Results:
We present CompleXChange, a tool for differential analysis of protein complexes based on predicted complexes and inferred complex abundances. For simulated data the results obtained by our complex abundance estimation algorithm are in better agreement with the ground truth and biologically more plausible than previous efforts that used linear programming. Also, execution time is much shorter.
The practical usability of the method was assessed in the context of transcription factor complexes predicted for human monocyte and lymphoblastoid samples. We demonstrate that our new method is robust against false-positive detection and reports deregulated complexomes that can only be partially explained by differential analysis of individual protein-coding genes. Furthermore we show that deregulated complexes identified by the tool potentially harbor significant yet unused information content.
Discussion:
CompleXChange yields orthogonal information to gene- and protein-centric analyses, which are not covering the physical interplay found in a cell. As no adequate experimental reference data exist to date, assessment of its performance was solely based on simulated data and biologic reasoning.

Protein vector representations for fast similarity search

Roman Feldbauer [(1)], Arthur Flexer [(1)], Thomas Rattei [(2)]
(1) Austrian Research Institute for Artificial Intelligence, Austria
(2) University of Vienna, Austria

Background:
Search in protein reference databases is a major bottleneck for projects involving large numbers of sequence reads.
This is due to hard to handle protein representations as well as fundamental problems of information retrieval from large databases.
We explore two complimentary approaches to accelerate similarity search:
(1) reducing computational cost of each single sequence comparison
(2) reducing computational complexity of search in databases
Methods and results:
(1) Distance measures in vector spaces are potentially much faster than sequence alignments or heuristics.
In order to measure protein similarity with Euclidean or Hamming distance proteins must be embedded in vector spaces with semantically meaningful geometry.
We investigate various embedding strategies, including FastMap projections, multidimensional scaling, and deep metric learning.
Preliminary results show that FastMap embeddings yield the correct best hit in many cases.
(2) Approximate nearest neighbor (ANN) methods allow for database queries in sublinear time.
We evaluate several methods, such as locality-sensitive hashing, product quantization, and specialized tree or graph structures on top of protein vector representations.
Preliminary results indicate tremendous speedups and only minimal loss in search accuracy.

Discussion:
Protein vector representations indexed in efficient ANN structures
could alleviate problems of ever-growing databases.
We present a proof-of-concept to this approach,
and discuss the ongoing and projected further development.
Major improvements are expected by integrating deep learning into the pipeline:
We will train specialized deep networks with alignment scores precomputed for the SIMAP2 database,
and thus forge vector spaces with geometries tuned to reflect trillions of exact sequence alignments.

M23D - MHC-II Model Database

Josef Laimer [(1)], Peter Lackner [(1)]
(1) University of Salzburg, Austria

MHC class II molecules are essential for initiating immune responses. Knowledge about their 3D structure is a potential key to understand their capabilities of binding antigen peptides or their interaction with T-cell receptors.
We present M23D, a database of 3D models of MHC-II molecules, currently for all HLA-DR alleles listed in the IPD-IMGT/HLA database. Its final version will contain data for all HLA class II alleles. M23D is indented for applications in antigen peptide binding studies or the analysis of fold stability.
The database provides experimentally determined structures, derived from the PDB database, as well as predicted models utilizing MODELLER. All models are scored with a set of widely used evaluation tools. Entries are cross-linked to source databases.
M23D is free for non-commercial usage and can be accessed via an easy to use web interface as well as a REST web service.

Tamock - Sample-specific gold standards in metagenomics

Samuel Gerner [(1)], Alexandra Graf [(2)], Thomas Rattei [(1)]
(1) University of Vienna, Austria
(2) FH Campus Wien, Austria

Background
In metagenomics, genome-centric approaches aim to retrieve original genomes of community members by assembling all sequences and separating (binning) the resulting contigs into genome bins using a range of discriminatory factors. These genome bins enable detailed insight into the metagenomic community on a genome level to study contributions and function of single microbial community members.
Such communities vary strongly in e.g. diversity and composition, leading to different method performances. Benchmarks like the CAMI Challenge (Sczyrba et al. 2017) use simulated data with known truths to assess recall and precision of respective methods to enable informed method selection. Nevertheless, pure simulated data cannot fully represent biological

conditions, especially if major fractions of the real data are unknown.
Methods and Results
To bypass this limitation, we developed tamock, a fully automated framework sampling all known sequences from reference databases while sequences of unknown origin are kept. This approach provides exact numbers and known truth of the known bacterial fraction while still maintaining original sample complexity. Using these gold standards, we could demonstrate the differing performance of separate assemblers and binning methods as expected.
Discussion
Resulting gold standards enable a study-specific method selection with minimal effort for benchmark creation. Unknown sequence fractions can heavily impact method performance as well as represent biological important data, making them an integral part for method selection for a specific study. Comparing results to analogues of real samples provides valuable insight for result interpretation as well as shortcomings and biases of respective methods.

Multilayered Network Framework for Biomedical Data Integration and Disease Gene Discovery

Pisanu Buphamalai [(1)], Tomislav Kokotovic [(2)], Vanja Nagy [(2)], Jörg Menche [(1)]
(1) CeMM, Austria
(2) Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases, Austria

Despite over a decade of post-genomic molecular biology, the functional characterisation of human genes in both healthy and disease states remains a challenge. Given the wealth of unbiased, high-throughput molecular and phenotypic data that is now available on many diseases, we can now attempt to go beyond studying one disease at a time and move towards a more systematic approach to identifying pathobiological processes in which the respective causal genes are involved. To represent the cross-scale nature of complex biological systems, we constructed multilayered networks derived from both physical and functional data, contributing to 22 layers of network. Mapping disease genes onto the different network layers reveals various disease localisation patterns that reflect important pathobiological mechanisms across the broad spectrum of human disease. This information can subsequently be incorporated into our disease gene prioritisation pipeline based on random walk with restart (RWR) algorithm, which we modified such that it takes network relevance of a disease as a guidance for network propagation. As an initial case study for such an approach, we consider a group of diseases characterised by intellectual disability (ID), where over 50% of causal genes are yet to be discovered. We found that informed propagation on disease-relevant network layers is able to retrieve left-out disease genes more accurately than using single layer or all layers without assigning relevant scores. On one hand, the results can facilitate further experimental validation and characterisation of newly identified genes. On the other hand, this framework can potentially contribute to fundamental understanding of complex genotype to phenotype relationship as well as serve the community as a platform for traceable biological data integration.

Metabolic adaptation of Chlamydia trachomatis during infection and RNA sequencing analysis

of host response in STI patients

Manli Yang [(1)], Karthika Rajeeve [(2)], Caroline Genco [(3)], Paola Massari [(3)], Thomas Rudel [(2)], Thomas Dandekar [(1)]
(1) Department of Bioinformatics, Biozentrum, University of Würzburg, Germany
(2) Department of Microbiology, Biozentrum, University of Würzburg, Germany
(3) Department of Immunology, Tufts University School of Medicine, Boston, MA, United States

Metabolic adaptation to the host cell is of vital importance for obligate intracellular pathogens such as Chlamydia trachomatis (Ct). Ct has a unique biphasic development cycle. This causes difficulties in genetic manipulation. Little is known about how the metabolic activity of Ct exactly evolves in the host cell and how host cells specifically respond to Ct's infection compared to the common pathogen-induced response. First, we generated a genome-scale metabolic model (GEM) and analysed pathway fluxes during different infection time points by flux balance analysis. We demonstrate specific metabolic adaptations for Ct elementary body (EB) and reticulate body (RB) state. Moreover, glutamate is probably a significant nitrogen and energy source for Ct. We also analysed whole transcriptome sequencing data from Neisseria gonorrhoeae (GC) infected STI patients to extract differential expressions of both host and GC under the condition of GC-Ct co-infection. We compared three different transcriptomic pipelines, Stringtie/Ballgown, Stringtie/DESeq2 and Cufflinks/Cuffdiff. According to the genes and transcripts with significantly differential expression in co-infection group, we found that Ct's infection probably may increase the expression of tumor protein 53 (TP53) and transferrin receptor (TfR). We hypothesize that Ct's infection is probably able to induce host cell ferroptosis with the cooperation of GC.

Cancer patient stratification by Greedy Symmetric Nonnegative Matrix Factorization

Oliver Müller-Stricker [(1)], Lars Kaderali [(1)]
(1) University Medicine Greifswald, Germany

According to the WHO, cancer is the second leading cause of death globally, and was responsible for 8.8 million deaths in 2015. One key challenge in cancer research is the stratification of patients into groups that have different clinical outcomes.
Amongst other algorithm families, nonnegative matrix factorization (NMF) has been proven useful for this purpose. However, most approaches are unable to analyze multiple related datasets concurrently instead of analyzing each one separately. Also, most NMF algorithms rely on standard multiplicative updates, which oftentimes show slow convergence. Here, we introduce a graph-based clustering approach which utilizes a newly developed NMF algorithm, Greedy Symmetric Nonnegative Matrix Factorization (GSNMF).
Our graph-based approach enables the concurrent analysis of information from multiple related datasets. Further, by our approach, factor matrix elements which promise a large objective value decrease are chosen for update more often, leading to an update frequency proportional to their final values. This means that matrix elements making only minor contributions to the result are updated less frequently, thus saving computationally intensive factor updates.
We apply our algorithm on a multi-relational breast cancer dataset consisting of patient mRNA

as well as miRNA expression data. First, patient affinity graphs are constructed from the input datasets. Then GSNMF is applied on the resulting graphs, yielding the patient clustering. We compare GSNMF to current state-of-the-art patient stratification algorithms and show that our newly developed algorithm compares favorably to these.

Simultaneous background correction and classification of infrared microscopic pixel spectra using deep stacked autoencoders

Arne Peter Raulf [(1)], Axel Mosig [(1)]
(1) Bioinformatics Group, Chair Biophysics Ruhr-University Bochum, Germany

Background
Spectral histopathology (SHP) utilizes infrared microscopy to characterize the disease status of histopathological tissue sections. Conventionally, each pixel spectrum is analyzed in a two-step procedure: first, background artifacts in the spectrum are removed following a physical model, called resonance Mie scattering. Subesequently, supervised/unsupervised learning approaches are applied to the corrected spectra. This two-step procedure comes at a high cost in terms of computation time, as the subtraction of the non-linear interference between the actual spectrum and the resonance Mie scattering artifacts involves an interative procedure which is applied individually to each of usually millions of pixel spectra wihtin a dataset. With the advent of new infrared microscopes which accelerate spectral image acquisition by two orders of magnitude through quantum cascade lasers, preprocessing time has thus become a major obstacle for applying SHP in practice.
Methods/Results
Based on the universal approximation theorem, we hypothesize that a deep neural network is capable of simultaneously learning preprocessing and classification. With confirming this hpyothesis, we are also showing the importance of a strong regularization effect through pretraining. While deep feed-forward neural networks fail to meet this assumption in practice, we show that stacked contractive Autoencoders (SCAE) for pretraining and a fully connected feed-forward neural network in the finetuning/ transfer learning part indeed facilitate reliable "non-stop” classification with a pixelwise accuracy of up to 93% compared to gold standard.
Discussion
Using the conventional two-step procedure as gold standard, our proposed networks perform favourable on uncorrected raw data. Besides a significant speed-up, our neural network approach also promises a systematic way of transfering models, for instance when sample acquisition changes from embedded tissue to fresh tissue.

CLIP-Explorer: A Galaxy-Pipeline for iCLIP and eCLIP data.

Florian Heyl [(1)], Daniel Maticzka [(1)], Rolf Backofen [(1)]
(1) Albert-Ludwigs-Universität Freiburg, Germany

RNA is a well-known polymeric molecule which plays a fundamental role in so many regulatory processes like splicing or translation, where RBPs (RNA-binding proteins) take part in. Thousands of RBPs have been found in human cells; some of which have been linked to

diseases encompassing various types of cancer. CLIP-Seq (crosslinking immunoprecipitation in combination with high-throughput sequencing) facilitated the analysis of RBPs but many protocols like iCLIP or eCLIP require specific tools and parameter sets to analyse the resulting data. These obstacles make the analysis hard to reproduce and demand a tight cooperation with a bioinformatician.
CLIP-Explorer resolves these issues by providing a workflow in Galaxy to guide the user to scrutinise iCLIP and eCLIP data, allowing for a user friendly interface to make quick changes to parameters and the pipeline itself. Consequently, CLIP-Explorer is flexible enough to be used for other CLIP-Seq datatypes as well (e.g., PAR-CLIP). The pipeline was tested on eCLIP data of RBFOX2 for which it identified known binding motifs and targets. Three different peak calling algorithms were tested, showing the flexibility of the pipeline and the importance of the parameter and method selection for the overall result.

High-throughput Immune Repertoire Sequencing Analysis using the Software ImmunExplorer

Julia Vetter [(1)], Stephan Winkler [(1)], Thomas Fraunhofer [(2)], Susanne Schaller [(1)]
(1) University of Applied Sciences Upper Austria, Campus Hagenberg, Austria
(2) University of Applied Sciences Upper Austria, Campus Hagenberg, Germany

Profiling of the immune repertoire based on next generation sequencing of the variable regions of B and T cell receptors has become a promising field in immunological research in various specialties (e.g. allergy, oncology, autoimmune diseases, and transplantation). Due to the complexity of the human adaptive immune repertoire and the partially unknown processes of the adaptive immune system in case of various diseases, it needs an extensive research to gain knowledge and to determine the behavior of the immune system in reference to specific diseases or in transplantation processes.
Therefore, we have designed and developed ImmunExplorer, a framework for analyzing B and T cell receptor data. ImmunExplorer enables the analysis of raw next generation data using the MiXCR tool, which has been integrated in ImmunExplorer. Furthermore, extensive clonality and diversity approaches are implemented as well as comparison and primer analyses. We also have designed a workflow for modelling the states of adaptive immune systems by analyzing B and T cell receptor repertoires using high throughput data. We currently use our workflow analyzing NGS data of blood and tissues from diseased patients, who have suffered from different kidney diseases as well as of blood from healthy patients. Features can be determined using the clonality and diversity of the immune repertoire sequencing data. Different machine learning algorithms have been integrated in ImmunExplorer for classifying and distinguishing between healthy and ill patient samples up to 84% accuracy.
ImmunExplorer 1.0 is currently available on http://bioinformatics.fh-hagenberg.at/immunexplorer/, and ImmunExplorer 2.0 will be published soon.

Identifying functionally interacting proteins via their phylogenetic profiles

Carla Mölbert [(1)], Ngoc-Vinh Tran [(1)], Ingo Ebersberger [(2)]
(1) Goethe University Frankfurt, Germany
(2) Goethe University Frankfurt, Senckenberg Biodiversity and Climate Research Centre, Germany

Background
The identification of proteins involved in the same biological process is an essential step in the annotation of newly sequenced genomes. Establishing orthology relationships to proteins organized in pathway databases, such as KEGG, is a common approach. However, this misses interactions that are confined to functionally poorly studied clades in the tree of life. The similarity/distance between the phylogenetic profiles for two genes can provide complementary information. We have added the option to cluster genes according to the distance of their phylogenetic profiles into our software PhyloProfile (https://github.com/BIONF/phyloprofile). Using yeast as a showcase, we survey if, and to what extent, phylogenetic profiles help to identify functionally interacting proteins.
Results
We have implemented four measures informing about the extent to which two phylogenetic profiles agree: (i) euclidean distance, (ii) Pearson's correlation coefficient, (iii) distance correlation, and (iv) mutual information. Phylogenetic profiles for the yeast gene set were determined with HaMStR-OneSeq (https://github.com/BIONF/hamstr), scoring for each yeast protein the feature architecture similarity (FAS) to each of the detected orthologs. This allows to optionally consider besides binary phylogenetic profiles, also non-binary profiles where the FAS score indicates the presence of an ortholog. Our initial benchmark reveals that all four measures perform comparably in clustering genes that are part of the same pathway or share a molecular function. Interestingly, physically interacting proteins show a lesser tendency to cluster.
Discussion
PhyloProfile can now be routinely applied to screen protein collections for interacting partners, even when their molecular function is unknown.

Amino acid residue pseudopotentials, SNP annotations using 3D-structures

André Marquardt [(1)], Tobias Juhre [(2)], Elisabeth Mack [(1)], Alexander Goesmann [(2)]
(1) Department of Hematology, Oncology and Immunology, University Hospital Giessen and Marburg, Germany
(2) Bioinformatics and Systems Biology, Gießen University, Germany

Cancer is caused by somatic mutations that change protein structure and function. Such mutations include single nucleotide polymorphisms (SNPs). Next Generation Sequencing (NGS) allows researchers to rapidly identify such mutations. Typically, previously unknown sequence variations are often detected. The interpretation of genetic mutations is currently limited to known data, available in database collections such as ClinVar and dbSNP. For classification of new mutations, many different algorithms exist, summarized in the dbNSFP. Unfortunately, these algorithms are rarely conclusive and currently only use secondary structure analyses.

Here we present a novel approach that includes the 3D structure of proteins that are affected by a sequence variation. We used amino acid residual pseudopotentials (APs) to transform the physicochemical properties of the protein into an energy measure, thus converting the 3D protein structure into a one-dimensional energy profile (EP).
Although the data basis allows only the consideration of some selected proteins, it was possible to correctly distinguish pathogenic from benign SNPS for 43 of 50 analysed mutations with a precision of 0.963 and an accuracy of 0.86, for globular proteins. Additionally, 13/13 SNPs were predicted correctly in a transmembrane protein.
Using APs offers a reliant way to classify SNPs, that may be used to identify novel driver mutations in cancer. The small data basis will presumably increase in the near future, based on the emerging use of cryo-electron microscopy. The approach presented here could then offer the possibility to quickly and reliably classify unknown mutations in patient samples leading to novel therapy options.

Gene regulatory networks in murine models of heart disease

Thiago Britto-Borges [(1)], Shirin Doroudgar [(1)], Mirko Völkers [(1)], Christoph Dieterich [(1)]
(1) University Hospital Heidelberg, Germany

Gene regulatory networks are central to the comprehension of human complex diseases, and the regulation of mRNA levels is a primary topic within this theme. The role of transcription factors (TFs) is particularly important since this type of regulators are known to enhance or repress transcription by associating with locus control regions within the genome, leading to differential gene expression (DGE) and hence been pivotal to the understanding of diseases. We here use a multi-omics approach, combining RNA-Seq DGE expression analysis with proteomic TF levels, to study how TF levels affect DGE in murine models for heart failure. Two models emulate pathological stress of disease, transverse aortic constriction and ischemia-reperfusion, while third model sedentary-swim represents physiological stress, totalising eight comparisons of conditions at different time points. Selecting genes called as differentially expressed in two of four methods - Cuffdiff, DESeq2, Limma and EdgeR - show only modest overlap of genes across conditions. We have then examined the enrichment of biological process terms with TopGO, demonstrating a higher degree of agreement across conditions. Surprisingly, this analysis shows a switch of disease-related terms between the physiological and pathological stress conditions. Next, we will use gene co-expression networks to create gene modules and determine whether specific TFs coordinate the observed switch between the conditions.
This work gives an overview of the differential regulatory mechanism on modules of genes in animal models for heart failure. We also plan to study post-transcriptional regulatory mechanisms, such as differential splicing, to extend our understanding of the transcriptional landscape in heart failure.

Landscape of RNA translation in murine models of heart disease

Etienne Boileau [(1)], Ilian Atanassov [(2)], Shirin Doroudgar [(1)], Mirko Völkers [(1)], Christoph Dieterich [(1)]
(1) Department of Internal Medicine III, University Hospital Heidelberg, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany
(2) Max Plank Institute for the Biology of Ageing, 50931, Koln, Germany

Background

Recent discoveries challenge our understanding of the coding potential of the genome. Amongst these are the small open reading frames (sORFs). Transcribed sORFs present a challenge for annotation, and so far little is known about their protein-coding capacity. Recent studies in a broad range of organisms have also revealed the translation potential of non-coding RNA as well as the importance of upstream open reading frames (uORFs) as a means of regulating translation [1]. uORFs are better known as cis-regulators of the translation of downstream canonical ORFs, but how this is modulated in vivo in specific tissues is yet to be characterised.

Ribosome profiling via high-throughput sequencing or Ribo-seq is a technique that allows to study translation at sub-codon resolution, revealing information about ribosome density and position along RNA transcripts [2]. When combined with affinity capture methods, this technique enables to profile specific cell populations in complex tissue contexts.

Methods and results

We used an extensive dataset of ribosome-profiling experiments, combined with a ribosome-tagging approach. This enabled us to characterise the translatome in mouse cardiac myocytes, in vivo, during stress conditions.

Various computational methods have been developed to classify and predict ORF translation based on Ribo-seq data. Using our software, Ribosome profiling with Bayesian predictions (Rp-Bp) [3], we were able to inventory alternative, small, upstream ORFs, annotated and predicted noncanonical translation events. A large number of predicted alternative proteins were also supported by proteomics evidence. We also revealed key insights into the extent and potential role of uORFs in the heart.

Discussion

In this work, we have given a comprehensive overview of the alternative translatome in the mouse heart. We were able to characterise translational control in cardiomyocytes in vivo in response to physiological and pathological stress. These results open the door to further functional proteins annotation with potential pathological importance.

[1] Couso, J. & Patraquim, P. Classification and function of small open reading frames Nat Rev Mol Cell Biol, 2017, 18, 575-589
[2] Aeschimann, F., Xiong, J., Arnold, A., Dieterich, C., & Grosshans, H. Transcriptome-wide measurement of ribosomal occupancy by ribosome profiling. Methods, 2015, 85, 75-89.

[3] Malone, B., Atanassov, I., Aeschimann, F., Li, X., Grosshans, H., & Dieterich, C. Bayesian prediction of rna translation from ribosome profiling. Nucleic Acids Res, 2017, 45, 2960-2972

Proteomics explains client specificity of the human translocon-associated protein complex in ER protein import

Duy Nguyen [(1)], Regine Stutz [(2)], Stefan Schorr [(2)], Sven Lang [(2)], Stefan Pfeffer [(3)], Hudson H. Freeze [(4)], Friedrich Förster [(5)], Volkhard Helms [(1)], Johanna Dudek [(2)], Richard Zimmermann [(2)]
(1) Center for Bioinformatics, Saarland University, Germany
(2) Medical Biochemistry and Molecular Biology, Saarland University, Germany
(3) Max-Planck Institute of Biochemistry, Department of Molecular Structural Biology, Germany
(4) Sanford-Burnham-Prebys Medical Discovery Institute, United States
(5) Bijvoet Center for Biomolecular Research, Utrecht University, Netherlands

In mammalian cells, one-third of all polypeptides are transported into or across the ER-membrane via the Sec61-channel. While the Sec61-complex facilitates translocation of all polypeptides with amino-terminal signal peptides (sp) or transmembrane helices (tmh), the Sec61-associated translocon-associated protein (TRAP)-complex supports translocation of only a subset, i.e. in a substrate-specific manner. To characterize TRAP dependent precursors, we combined siRNA-mediated TRAP depletion in HeLa cells, label-free quantitative proteomics, and differential protein abundance analysis. The results were validated in independent experiments by western blotting, quantitative RT-PCR, and complementation analysis. Sp analysis of TRAP-clients revealed above-average glycine-plus-proline content and below-average hydrophobicity as the distinguishing features. Thus, TRAP, which is linked to congenital disorders of glycosylation, may act as sp-receptor on the ER-membrane's cytosolic face for precursor polypeptides with high glycine-plus-proline content and/or low hydrophobicity in their sp and trigger substrate-specific (and regulated) opening of the Sec61-channel by interaction with the ER-lumenal hinge of Sec61α

Designing functional molecules with DNArchitect

Stephan Pabinger [(1)], Yasaman Ahmadi [(1)], Regina Soldo [(1)], Pia Glaser [(1)], Klemens Vierlinger [(1)], Ivan Barisic [(1)]
(1) AIT Austrian Institute of Technology, Austria

Background
Aptamers are oligonucleotides that bind to a specific target molecule. They can be used for both basic research and clinical purposes as macromolecular drugs. By combining them with a catalytic DNAzyme, these Aptamer-DNAzyme molecules have additional research and industrial applications as their catalytic activity can be activated in the presence/absence of the cognate ligand. The current challenge is to design the functional connection of ligand-specific aptamers and catalytic DNAzymes by using linker sequences, resulting in a molecule which unfolds itself in the presence of the specific aptamer ligand, causing a detectable catalytic activity.

Methods and results
We have created DNArchitect, a novel web-based tool, that facilitates designing these functional Aptamer-DNAzyme molecules. DNArchitect includes two methods [1], each supporting a distinctive design of the Aptamer-DNAzyme molecule, to calculate the linker sequence given an aptamer and a DNAzyme sequence. The included database allows users to store their aptamer and DNAzyme sequences and submit them for calculation. Both methods are highly customizable, run in the background, and store a list of possible DNA sequences in the database. The designed sequences are scored and ranked, can be exported for further analyses, and their calculated structure can be visualized.

Discussion
We have designed a novel tool to easily design Aptamer-DNAzyme molecules without bioinformatics requirements. The software is freely available and may help to reduce the burden of manually designing the Aptamer-DNAzyme sequences leading to better functional structures. DNArchitect can be accessed at http://tools.ait.ac.at/dnarchitect.

[1] http://2015.igem.org/Team:Heidelberg/software/jaws

White Box Modeling, Variable Impact Analysis, and Interaction Network Identification in Biological and Medical Data Using Symbolic Regression

Stephan Winkler [(1)], Susanne Schaller [(1)], Viktoria Dorfer [(1)], Andreas Haghofer [(2)], Julia Vetter [(2)], Michael Affenzeller [(3)]
(1) University of Applied Sciences Upper Austria; Hagenberg; BIN, HEAL, Austria
(2) University of Applied Sciences Upper Austria; Hagenberg; BIN, Austria
(3) University of Applied Sciences Upper Austria; Hagenberg; HEAL, BIN, Austria

The main goal of regression in general is to determine the relationship of a dependent (target) variable t to a set of specified independent (input) variables x. Thus, our aim is to identify a function f that uses x such that t = f(x).
Symbolic regression is the induction of mathematical expressions on data. The key feature of this approach is that the object of search is a symbolic description of a model. This is in sharp contrast with other methods of nonlinear regression, where a specific model is assumed. Using a set of basic functions we apply genetic programming as search technique that evolves symbolic regression models as combinations of basic functions through an evolutionary process. As the complexity of these models is limited, the process is forced to include only essentially relevant variables in the models; multi-criterial optimization can be used to guide the search process towards compacter, more concise models. This information reveals variable interactions, which can be shown as interaction models.
We have used this technique successfully in the analysis of medical and biological data. For example, we have so identified virtual tumor markers and tumor prediction models, prediction models for critical states in medical workflows, variable interactions of substances in apples, impacts of weather and process parameters on the microbial contamination of herbs, classification models for cells, validation models for peptide identifications, and models that classify the state of human adaptive immune systems on the basis of NGS data.

Discriminating Spontaneous Tumors from Exposure Induced Tumors in A/J Mice Lung Cancer Model

Yang Xiang [(1)], Florian Martin [(2)], Karsta Luettich [(2)], Keyur Trivedi [(2)], Emmanuel Guedj [(2)], Ee-Tsin Wong [(3)], Julia Hoeng [(2)], Manuel Peitsch [(2)]
(1) Philip Morris International, Switzerland
(2) Philip Morris International R&D, Switzerland
(3) Philip Morris International R&D, Singapore

Background
The A/J mouse model is highly susceptible to chemical lung tumor induction and has been widely used as a screening model in carcinogenicity testing and chemoprevention studies. Although cigarette smoke exposure induces tumors in the lungs, non-exposed A/J mice will also develop lung tumors spontaneously as they age. This raises the question whether the exposure-induced tumors are of a similar type to spontaneous tumors, irrespective of the overall exposure effect. As exposed mice may exhibit both exposure-induced and spontaneous tumors, this leads to a one-class problem as only spontaneous tumors arising from non-exposed animals are unequivocally defined. We have developed and applyed a one-class classifier to examine the potential differences between tumors developing in exposed vs unexposed A/J mice.
Methods and results
Based on lung gene expression profiles, genes having a specific behavior in spontaneous tumors as compared to tumors from exposed mice were ranked based on the statistical interaction model described in our previous publication (PMID: 26109882). A one-class classifier based on a penalized Mahalanobis distance was developed using the genes with the biggest absolute interaction values from this model. The performance was assessed by cross validation on the spontaneous tumor group, leading to correct classification in 75% of cases. Tumors arising in exposed animals are significantly distanced from the spontaneous tumors (t-test, p-value < 0.001). The obtained classifier was then applied to the lung gene expression data of mice exposed to THS2.2 aerosol, which is an erosol generated by a product developed by Philip Morris International aiming at reducing the risk of smoking-related diseases. The results showed that lung tumors of mice exposed to THS2.2 aerosol were significantly distanced from those of the cigarette smoke-exposed mice (p-value < 0.001) while being closer to spontaneous tumors.
Discussion
In this study, the feature selection bias has been addressed by leveraging an independant study. The selected genes also exhibited an high interaction term in this dataset, therefore the classifier is expected to be robust. The results of this study are promising and highlight the value of one-class classifiers in the instance where tumour types can not be easily characterized.

circtools: a one-stop software solution for circular RNA research

Tobias Jakobi [(1)], Alexey Uvarovskii [(1)], Christoph Dieterich [(1)]
(1) University Hospital Heidelberg, Germany

Background
Circular RNAs (circRNAs) originate through back-splicing events from linear primary transcripts, are resistant to exonucleases, not polyadenylated, and have been shown to be specific for cell type and developmental stage. However, for the majority of circRNAs, their function is yet to be determined. Prediction of circRNAs is a multi-stage bioinformatics process yielding, depending on tissue and condition, sets of hundreds of potential circRNA candidates that require further analyses. While a number of tools for the prediction process exist, available downstream tools are rare. We aim to provide researches with a harmonized work flow that covers different stages of in silico analyses, from prediction to first functional insights.
Methods and results
Here, we present circtools (http://circ.tools), a modular, Python-based framework that unifies several in silico circRNA analyses in a single command line-driven toolbox. Circtools includes modules for circRNA detection and reconstruction that are based on our well tested DCC (Cheng et al. 2016) and FUCHS (Metge et al. 2017) tools and was developed as a complete analysis work flow that contains additional modules to perform initial quality checks, test circRNAs for host gene independent expression, identify differentially spliced exons, screen circRNAs for enriched features, and design circRNA-specific primers for qRT-PCR verification.
Discussion
Circtools supports researchers with a complete circRNA work flow, complemented by visualization options and data export into commonly used formats. We intend to add more modules in the future in order to provide a comprehensive bioinformatics toolbox for the research community and encourage users to contribute new modules.

Multi-level network analysis of structural ensembles

Markus Schneider [(1)], Iris Antes [(1)]
(1) Technical University of Munich, Germany

Background
Protein function arises from structure and dynamics. The latter can be represented by time-ordered ensembles of structures as generated by molecular dynamics (MD). For a detailed understanding of many biological phenomena such as allostery, it is essential to analyse the structural changes during these processes. A network-based model of the corresponding time-ordered ensembles allows for intuitive visualization and analysis of the underlying structural features. Although a plethora of analysis tools exists for individual protein structures, there is a need for computational tools that can transform data from structural ensembles into networks.
Methods and results
We developed a plugin for the network visualization software Cytoscape, which reads and analyses interaction data extracted from MD-trajectories. The network can be explored on

multiple structural levels, from residues to atoms, using timeline averages or single time frames. We provide various analysis functions based on timeline properties. Multiple data files can be combined in order to allow simultaneous analysis of different interaction types such as H-bonds or hydrophobic contacts. For structural visualisation of selected network regions, the plugin can connect to the structure viewers PyMOL, Chimera or VMD.
Discussion
In comparison to similar tools focusing on static structures, our plugin imports ensemble timeline data and offers specialized analysis methods. It allows for dynamic switching between resolution and time levels, opening up new possibilities for exploration. Finally, our data import approach offers more flexibility by allowing free combinations of input files and providing a simple input format.

Genome and transcriptome characterization of glycoengineered Nicotiana benthamiana

Matteo Schiavinato [(1)], Richard Strasser [(2)], Lukas Mach [(2)], Juliane C. Dohm [(1)], Heinz Himmelbauer [(1)]
(1) University of Natural Resources and Life Sciences (BOKU), Department of Biotechnology, Vienna, Austria
(2) University of Natural Resources and Life Sciences (BOKU), Department of Applied Genetics and Cell Biology, Vienna, Austria

Many transgenic Nicotiana benthamiana lines have been generated to enable in planta production of glycoproteins, among which ΔXT/FT (Strasser et al. 2008). In this line, two families of glycosylation genes are knocked down through RNA interference. An analysis of its genome and transcriptome is addressed here. This line is one of many that are maintained by different institutions; however, little is known about their origin and genetic proximity to each other (Goodin et al. 2008). Here we also compare the transcriptomes of wild-type N. benthamiana lines used by different laboratories.
We performed a gene prediction on a genome assembly producing 50,516 genes (62,216 isoforms), of which 71% were then functionally annotated. We used this gene set to analyse the genome and transcriptome of ΔXT/FT, detecting differentially expressed genes with the original wild type. We used genomic paired-end reads to identify the insertion sites of the transgenes used in the plant, and we show the presence of a fusion transcript with a host gene. Lastly, we show that little variation can be detected between different research N. benthamiana lines, although not as low as for plants from the same accession.
With our study we make a step forward in characterizing the genome and transcriptome of N. benthamiana using ΔXT/FT as a template, and we show the importance of such analysis when using a transgenic plant. We also prove that a closely related group of accessions is likely to be at the origin of the N. benthamiana lines used in research today.

Towards physicochemical bioinformatics: an interactive web-based tool for visualizing and comparing physicochemical properties of biological sequences

Lukas Bartonek [(1)], Bojan Zagrovic [(1)]
(1) University of Vienna, Austria

Variation in the physicochemical properties along the primary sequence of macromolecules is often visualized in the form of one-dimensional profiles - line graphs capturing the value of a physicochemical property of interest (charge, hydrophobicity, disorder etc.) against sequence position. Such representation is frequently used by researchers to compare sequences at a glance without utilizing an in-depth analysis and is usually implemented on a case-by-case basis in a static manner. The use of static visualization methods, however, can lead to misinterpretation of the underlying data due to fixed visualization parameters.
We present a server that allows users to easily create interactive and flexible visualizations of biomolecular sequence properties and explore the data using multiple parameters. Both protein and RNA sequences are supported and several property scales are provided with an option of easily adding other ones. By performing most calculations client-side utilizing a Javascript interface and D3.js for data visualization, the server enables real-time interactive sessions. Users can compare multiple sequences, each with potentially different physicochemical properties visualized, with a correlation score dynamically updated to provide a measure of similarity.
A high degree of interactivity enabled by our tool improves the understanding of the underlying data. The ability to tune visualization parameters, including the smoothing method used, and shift sequences against each other, enables a rapid and multifaceted exploration of data sets of interest. By treating biomolecular sequences as physicochemical objects, which they invariably are, our tool enables detection of biologically meaningful similarities between sequences that may be inaccessible to standard methods of primary sequence comparison.

Processing of paper-based case report forms using optical character recognition software for automated data transfer

Sabine Hurka [(1)], Ruth Ajnwojner [(1)], Christian Troidl [(1)], Oliver Dörr [(2)], Holger Nef [(2)], Christian W. Hamm [(3)], Christoph Liebetrau [(1)], Till Keller [(1)]
(1) Department of Cardiology, Kerckhoff Heart and Thorax Center, Germany
(2) Department of Cardiology, University of Giessen, Germany
(3) Department of Cardiology, Kerckhoff Heart and Thorax Center; Department of Cardiology, University of Giessen, Germany

Background:
Successful execution of clinical studies strongly depends on the quality of the collected data, including data assessment and data management. We aimed to establish a process of automated data capture of paper-based case report forms (CRFs) within the framework of a biomarker registry that enrolls more than 100 new participants per month.
Methods and results:
Clinical data is stored using a non-profit software (REDCap) on a server hosted at our institution. Within this project, CRFs were developed based on variables used in clinical trials of the German Centre for Cardiovascular Research (DZHK). These paper-based CRFs are

filled out at the bedside, scanned, and archived as PDF by study staff. We adapted a commercial optical character recognition (OCR) software package (ABBYY FlexiCapture 11) to transfer the data into our database. Also, we used JavaScript during OCR for customization and data validation.
The accuracy of OCR depended strongly on how carefully the CRFs were filled out. The mean hands-on time for onscreen checking and correction of the data entries for 50 CRFs was 32.03 +/- 7.3s per individual CRF. Here, 3.37% of the parameters had to be corrected manually. This included 0.93% human errors such as poorly formed letters and 2.44% errors in cases the OCR software decided wrong.
Discussion:
We successfully implemented an OCR-based semi-automated system for data processing of paper-based CRFs and for data transfer into a study database. In the next phase we plan to develop this process further so that it is fully automated.

Assessment of open reading frame coding potential using proteomic and genomic data.

Anastasiia Padalko [(1)], Thomas Rattei [(1)], Harald Marx [(1)]
(1) University of Vienna, Germany

Background: Differentiating between coding and non-coding open reading frames (ORFs) is a fundamental step to the comprehensive analysis of biological systems. Conventional ORF classification approaches rely on genomic and transcriptomic features. However, most classifiers are not designed to handle small ORFs (smORFs) due to an inherent length cutoff (≤ 100 codons) and missing genetic signals, e.g. non-canonical start codons. The continuous advancement of mass spectrometry(MS)-based proteomics results in generation of complementary high-throughput data to next generation sequencing data, that enables significant improvement of current ORF coding potential assessment methods. Methods: The project is focused on developing a Java pipeline which enables classification of both canonical and small ORFs. In order to accomplish the objective, we implement a machine-learning approach that combines genomic and proteomic features. The proteomic features are retrieved from large-scale shotgun proteomics experiments (PRIDE repository). Well-annotated human smORFs (sORF.org), non-coding ORFs, and canonical ORFs (Ensembl) serve as search database and respective training set. The pipeline is compared against state-of-the-art classifiers. Discussion: Biological significance of smORF-encoded polypeptides (SEPs) highlights the importance of implementing efficient computational tools for their classification. Incorporation of the proteomic features provides an extra layer of evidence for ORF discrimination independent of gene structure heterogeneity and conservation levels.

From Minimum Inhibitory Concentrations to Resistance Phenotypes: Semi-Automated Breakpoint Interpretation

Victoria Beneder [(1)], Peter Sykacek [(1)], Stephan Beisken [(2)], Andreas Posch [(2)]
(1) Department of Biotechnology, University of Natural Resources and Life Sciences, Austria
(2) Ares Genetics GmbH, Austria

Background
Antimicrobial resistance is a serious threat to global health, resulting in high medical costs and increased mortality. For severe infections, a pathogen's resistance phenotype may be needed for treatment decisions. The European Committee on Antimicrobial Susceptibility Testing (EUCAST) regularly publishes guidelines on how to derive resistance phenotypes from antimicrobial susceptibility testing (AST) results.
Designed for human use, they contain different table structures and free text, posing a challenge for automated data analysis. Additionally, guidelines are defined for different taxonomic ranks such as species or families, or phenotypically-defined groups, such as coagulase-negative staphylococci.
Our tool unifies information encoded in EUCAST's clinical breakpoints, expert rules and intrinsic resistances publications into an adaptable in silico system, enabling resistance phenotypes lookup for species-compound combinations and sample classification as resistant or susceptible if minimum inhibitory concentrations are provided. Intrinsic resistances and interpretive rules via resistance phenotypes add additional resistance information.
Methods and Result
We use the Python environment for data extraction, converting different data structures of the EUCAST publications to a standardized format and scanning the associated free texts for restrictions in the application of the breakpoints. To consider relevant breakpoints defined on any taxonomic rank, we map all organisms against the NCBI taxonomy database and replace phenotypically defined groups by the species they contain. We use a rule based engine, encoding resistance phenotype patterns that indicate the presence of other linked resistances in the tested sample to derive further resistances phenotypes.
Our tool compiles all EUCAST guidelines into a uniform system that translates AST results to resistance phenotypes and derives information from intrinsic resistances and interpretative rules, leading to programmatic analysis and easier access to AST interpretation.

Live analysis and privacy-preserving real-time filtering in next-generation sequencing while the sequencer is still running (Poster to the talk)

Tobias Loka [(1)], Simon H. Tausch [(1)], Martin S. Lindner [(1)], Piotr Wojtek Dabrowski [(1)], Benjamin Strauch [(1)], Jakob M. Schulze [(1)], Aleksandar Radonić [(1)], Andreas Nitsche [(1)], Bernhard Renard [(1)]

(1) Robert Koch Institute, Berlin, Germany

Background: With the continuously increased use of next-generation sequencing (NGS) in time-critical applications such as disease outbreak analysis and precision medicine, there is a strong need for fast turnaround time from sample arrival to analysis results. At the same time, sequencing data in these fields may contain sensitive information that enable the re-identification of human individuals and other privacy-breaching strategies even for anonymized data. Thus, powerful methods for the analysis of NGS data that provide early interpretable results are required while also taking data protection into account.
Methods and results: We developed a collection of tools for the analysis of NGS data while the sequencer is still running. This includes software for read mapping (HiLive; Lindner et al., 2017, doi:10.1093/bioinformatics/btw659), taxonomic classification (LiveKraken; Tausch et al., 2018, doi:10.1093/bioinformatics/bty433) and pathogen identification (PathoLive). PriLive is a novel tool for the automated removal of sensitive data while the sequencing machine is

running (Loka et al., 2018, doi:10.1093/bioinformatics/bty128). Thereby, reads related to a specified set of organisms of interest are unaffected. With a sensitivity of >99.4% and a specificity of >99.8% even for reads with a high number of variations, PriLive achieves results at least as accurate as conventional post-hoc filtering tools.
Discussion: With PriLive, we implemented a solution for an increased level of data protection by removing human sequence information before being completely produced. This strongly facilitates the compliance with strict data protection regulations and simplifies subsequent analyses of sensitive data.

Machine-Learning-Based Annotation Triage and Extraction of Microbial Phenotypic Traits from the Scientific Literature

Lukas Lüftinger [(1)], Thomas Rattei [(1)]
(1) Department of Computational Systems Biology, University of Vienna, Austria

Introduction
As the growth rate of the biomedical literature continues to rise, the need for automatic information extraction from scientific texts likewise increases. I thus investigate the possibility of using supervised machine learning to select scientific articles for manual annotation of bacterial phenotypic traits. Additionally, I attempt the fully automatic extraction and downstream application of trait information from full text.
Methods and Results
A Python framework was developed for text extraction, curation and efficient manual annotation of scientific articles for the presence and sign of phenotypic traits. A set of support vector machine (SVM) and convolutional neural network (CNN) classifiers are trained on such a manually annotated data set, and used to predict traits in the literature. Phenotypic trait information is then used to label bacterial reference genomes for the PhenDB workflow. The PhenDB pipeline utilizes machine learning models trained on labeled functional representations of microbial genomes, and allows for the rapid prediction of functions and traits from unlabeled genomes, for example from a metagenomics experiment.
Several PhenDB models trained with text mined genome labels significantly out-performed pre-existing models.
Discussion
While the extraction of arbitrary phenotypic traits and from arbitrary scientific article types remains challenging, the method yields good results on certain subsets of either. Combining pure machine learning methods with rule based approaches can, while reducing method flexibility, further increase prediction accuracy.

Network-based approach to drug-gene interactions

Loan Vulliard [(1)], Jörg Menche [(1)]
(1) Research Center for Molecular Medicine of the Austrian Academy of Sciences, Austria

Being able to stratify individuals according to whether a given drug will provide an adapted treatment to their condition or not is highly challenging and requires a deep understanding of

the rules underlying drug effectiveness. This project is motivated by the hypothesis that some biological principles are underlying a cell's response to perturbations, and by studying how intrinsic and extrinsic cell perturbations combine and affect the cell morphology with a systematic approach some of these rules could be identified.
To address these challenges at the molecular level, we will perform and analyze an arrayed morphological screen, comparing genetic and chemical perturbations in MCF10A, a human epithelial cell line, using small molecules and CRISPR knockouts separately and in combination. High Content Imaging will allow us to extract the corresponding comprehensive morphological profiles. Our group has recently developed a methodology providing detailed interaction maps from such high-dimensional morphological data, giving a precise assessment of how exactly genes and drugs interact with each other. The results can then be represented as a multi-edged bipartite network with genes on one side and drugs on the other, and easily interpreted and integrated with external annotations from public databases. This poster will present the results obtained so far while setting up the screening process, highlight how a computational approach was used in the conceptualization and planning of the project and underline how machine learning and networks science can help from experimental design to the analysis of the results.

NICE – Network Informed funCtional Enrichment

Felix Müller [(1)], Michael Caldera [(1)], Sebastian Pirch [(1)], Jörg Menche [(1)]
(1) Research Center for Molecular Medicine (CeMM), Austria

The integrated network of all physical protein interactions within the cell can be interpreted as a map of biological mechanisms.
Functional annotations of genes together with their neighbors can be used to generate functional landscapes of such biological networks.
We develop a tool called NICE (Network Informed funCtional Enrichment) that characterizes biological networks by identifying functional similarities
based on GO terms, disease- or phenotype associations etc. Dimension reduction techniques such as t-SNE allow us to generate specific
network layouts that helps us study and predict relationships among genes.
We explore functional landscapes on the interactome that serve as cross-reference data to sample gene sets from experiments.
Additionally, we develop a virtual reality platform called NetDiVR to interact with big datasets such as large networks in virtual 3D space,
that enables us to combine the analytical power of computer algorithms with the intuition and creativity of researchers.

Benchmarking the impact of long-read correction methods

Philipp Peters [(1)], Nancy Stralis Pavese [(1)], Heinz Himmelbauer [(1)], Juliane C. Dohm [(1)]
(1) University of Natural Resources and Life Sciences (BOKU), Department of Biotechnology, Vienna, Austria, Austria

Accurately assembled genomes are crucial in order to gain insights into genome structure and evolution of an organism. To achieve such an assembly for genome-wide analyses a variety of sequencing technologies can be used that differ in length and error rate of the generated raw data. Long reads can especially serve to improve the contiguity of genome assemblies and facilitate the detection of structural variants between genomes. Currently, sequencing devices by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are able to produce such reads spanning several kilobases. However, reads from these platforms comprise error rates of approximately 15%, hence have to be corrected during the assembly process. A specific challenge for the correction step is the device-dependent difference in the distribution of errors in a read. Furthermore, the paradigms and the correction methods themselves differ between pipelines. The main approaches are 1) consensus correction either with a subset of the long-read data or with additional less erroneous short reads, and 2) patching of larger fragments in order to keep longer reads. Different correction steps lead to differing error rates and lengths of the corrected reads. In this poster, we investigate the specific error rate of reads from different platforms and benchmark different correction strategies.

Detection of haplotype blocks for genome scaffolding in plants

Alexandrina Bodrug [(1)], Hermann Bürstmayr [(1)], Nancy Stralis-Pavese [(1)], Juliane C. Dohm [(2)], Heinz Himmelbauer [(1)]
(1) University of Natural Resources and Life Sciences Vienna, Austria
(2) University of Natural Resources and Life Sciences (BOKU), Austria

Whole genome resequencing data of representative individuals of a species is a resource generated for population genomics studies. We propose to use this resource for scaffolding, one of the last steps of genome assembly and a long-lasting challenge when improving contiguity of non-reference genomes. Genomic regions can be characterized by a set of haplotypes in a population. We start from the biological observation that a collection of individuals is going to carry blocks of similar haplotypes throughout their genomes because of a shared ancestry. We use single nucleotide polymorphisms from variant calling datasets of plant genomes to build haplotype blocks, then use this information to create connections based on the re-appearance of the same set of blocks in multiple regions. We find that provided a sufficient number of individuals and after exclusion of uninformative positions we can reconstruct the architecture of known genomes and consequently scaffold fragmented genome assemblies. The method was applied to Arabidopsis thaliana and to Chenopodium quinoa. It is implemented in perl and uses string distance to build haplotype blocks and detect connected regions. The success of the scaffolding based on haplotypes will vary for different species and for different datasets. This is caused by the fact that the number and similarity of haplotypes in a genomic region are dependent on the size and heterogeneity of the population used for variant calling as well as the genetic history of the species.

Application of Machine Learning and Computer Vision Methods to Characterize Cancer Cells in Hodgkin Lymphoma

Ben Haladik [(1)], Hendrik Schäfer [(2)], Sylvia Hartmann [(3)], Martin-Leo Hansmann [(3)], Ina Koch [(1)]
(1) Goethe University Frankfurt, Germany
(2) Hansmann Lieberz GmbH, Germany
(3) University Hospital Frankfurt, Germany

Background
Classical Hodgkin lymphoma is a malignancy of the lymphatic system [1]. Its malignant cells always express the marker CD30 and show a very peculiar morphology. For diagnosis, these cells are often counted and characterized manually. Therefore, an automated quantitative analysis of tumor cells and their nuclei is of great interest for cancer research. Due to the size of whole slide images (WSI), difficulties in segmenting cells in tissues and different kinds of artifacts in WSI, these analyses are cumbersome.

Methods and Results
We analyzed 46 double-stained CD30/hematoxylin WSI of first and recurrent instances of Hodgkin lymphoma. Using Support Vector Machines, we can detect tissue for further segmentation with over 99% accuracy and measure tissue sizes. To detect CD30+ cells, we use a Convolutional Neural Network with the U-Net architecture [2]. A pixel-wise categorical accuracy of 93% on the validation dataset is achieved by explicitly considering different kinds of artifacts. We use this network in a pipeline to characterize the spatial organization of CD30+ cells in tissue. Filtering and Sauvola thresholding [3] is used to segment nuclei of CD30+ cells. Subsequently, we analyze the morphology of these nuclei.
Discussion
With object-wise precision values between 64% and 87% per image and recall values between 73% and 100% per image, our method performs similarly to other state-of-the-art methods for segmentation of immunohistological images. We discuss differences in the spatial distribution and morphology in recurrent and new instances of Hodgkin Lymphoma. However, the reliability of morphological characterizations strongly depends on the segmentation performance.
[1] Küppers et al., The Journal of Clinical Investigation, 122:3439, 2012
[2] Ronneberger et al., International Conference on Medical Image Computing and Computer-assisted Intervention, 234, 2015
[3] Sauvola et al., Pattern Recognition, 33:2, 2000

Modelling Segmental Duplications in the Human Genome

Eldar Abdullaev [(1)], Peter F Arndt [(1)]
(1) Max Planck Institute for Molecular Genetics, Germany

Segmental duplications (SDs) are long (> 1kbp) DNA sequences that are repeated (sometimes multiple times) in a genome and have high sequence identity. There are several

well-studied mechanisms that are responsible for propagation of segmental duplications: non-homologous end joining, DNA polymerase slippage, non-allelic homologous recombination etc. However, we do not have a general understanding that would explain the distribution of SDs in the genome, for example: why do SDs appear more often in some regions than in others, why do SDs overlap with each other so often, how does selection affect their distribution in the genome etc. Up to now there is no mathematical model for the propagation of SDs proposed that would explain simple statistical features of SDs. The goal of our project is to find such a model. We use a graph representation of all SDs in the human genome and try to approach this system from a complex network point of view. Nodes and edges of the graph represent genomic regions and homology between two regions (alignment), respectively. The SD graph appears similar to those of scale-free complex networks and some graph features shed light on the dynamics of the propagation process.

Determine ambiguous regions in public genomes

Calin Rares Lucaciu [(1)], Thomas Rattei [(1)]
(1) University of Vienna, Austria

Sequenced genomes present in public databases are a valuable resource for species characterization. Due to the risk of contamination in the process of sample preparation and errors generated from sequencing technology, the quality of these genomes is often questioned. Methods for finding contaminated regions in the genomes have been proposed which mostly relies on prediction based on structure composition or similarity to other species. These methods are limited by the databases and by the lack of knowledge about evolutionary aspects.
Our proposed method for finding ambiguous-regions in the genomes is based on coverage information calculated from short-reads alignment. It relies on the assumption that the coverage follows a normal distribution over the entire genome. The mean and standard deviation of the distribution are calculated from the mean-coverage of randomly selected genomic fragments. By using a sliding-window for every genomic fragment the mean-coverage is calculated, and ambiguous regions are determined based on the defined z-score threshold. The method was tested so far on several Arabidopsis genomes and such ambiguous regions genomes were found but not further tested yet.
Ambiguous covered regions may be indicative for contamination present in the genome that might have led to wrongly-assembled genomes. The method might be validated by removing from the data sequences that aligned to such regions or by re-sequencing a new sample from same species and reassemble the genome.

GENOME-WIDE SCAN FOR DIAGNOSTIC MARKERS FOR BUD BURST IN BEECH

Malte Mader [(1)], Niels Andreas Müller [(1)], Matthias Fladung [(1)], Bernd Degen [(1)], Mirko Liesebach [(1)], Heike Liesebach [(1)], Birgit Kersten [(1)]
(1) Thünen Institute of Forest Genetics, Germany

Background

Bud burst in trees marks the transition from a dormant to an active phase and is vulnerable to adverse conditions. Late spring frosts may damage beech trees if the bud burst occurs too early in the season. Late bud burst, in contrast, can result in an incomplete utilization of the growing period. A high heritability of this important adaptive trait in different tree species indicates an involvement of genetic factors. Here, we aim to identify diagnostic genetic markers for early bud burst in beech using a genome-wide approach in the scope of the GenMon project (https://www.gen-mon.de/).
Methods and results
To develop diagnostic markers for bud burst, one tree with early and one tree with late bud burst were selected from each of 14 different provenances. DNA from the 14 early trees and the 14 late trees was pooled and sequenced (HiSeq; 2x150bp, 84x). After mapping the pool data to a recently published reference genome sequence of beech (Mishra et al, 2018) we identified 6.8 million SNPs that passed our filtering criteria using the R programming language. A genome scan using a sliding window approach over all SNPs on all scaffolds revealed several genomic regions exhibiting exceptionally high numbers of SNPs with distinct allele frequency differences between the two pools.
Discussion
These regions may contain the allelic variants responsible for the differences in bud burst and thus serve for the development of diagnostic markers. Potential markers will be validated in an extended set of phenotyped beech trees.

References
Bagdevi Mishra, Deepak K Gupta, Markus Pfenninger, Thomas Hickler, Ewald Langer, Bora Nam, Juraj Paule, Rahul Sharma, Bartosz Ulaszewski, Joanna Warmbier, Jaroslaw Burczyk, Marco Thines; (2018) A reference genome of the European beech (Fagus sylvatica L.), GigaScience, Volume 7, Issue 6

The revised and extended BRENDA Structure Search

Lisa Jeske [(1)], Dietmar Schomburg [(1)]
(1) Dpt. for Bioinformatics and Biochemistry, BRICS, Technische Universität Braunschweig, Germany

Efficient searching for molecular structures is a challenging task in cheminformatics. Structure searches are often used in cases where a structure can be proposed for an unknown compound or as a screening method for drug discovery like the development of novel enzyme inhibitors. In BRENDA, it is possible to search chemical structures either conventionally by name or by drawing by means of a chemical structure editor of ~133,000 different molecules in the ligand database containing substrates, products, inhibitors, activating compounds, and cofactors. In addition to the former substructure search (1), two new search types are provided.
The new similarity search uses a fingerprint scan taking into account the Tanimoto coefficient - a measure for similarity - to detect suitable molecules. A fingerprint of a molecule is based on a set of all paths of atoms and bonds within the molecule with a maximal length of 8 atoms.
In the implemented isomer search, a complete graph matching algorithm is executed to compare the drawn structure with the BRENDA ligand structure. The fingerprint scan as a

prefilter is carried out before the full structure search to reduce the runtime and to decrease the number of time-consuming graph matchings.
A new progress bar informs the user about the current status of the program. The search form and result pages have been adapted to improve the user-friendliness and clarity.
Furthermore, the ligands on the result page can now be filtered by organism and EC number.
References
1. Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res., 32, 431D–433.

NoPeak: Binding Motif Discovery from ChIP-Seq Data without Peak Calling

Michael Menzel [(1)], Andreas Gogol-Döring [(1)], Sabine Hurka [(2)]
(1) Technische Hochschule Mittelhessen, Germany
(2) Justus-Liebig-Universität Gießen, Germany

Background
Binding motif discovery through Chromatin ImmunoPrecipitation with high-throughput DNA Sequencing (ChIP-Seq) data is an important tool for understanding regulatory processes. Identifying binding motifs from ChIP-Seq data is usually based on peak calling where regions with an enriched read count are classified as peaks and subsequently the surrounding regions are analyzed using a motif finder. Yet, only part of the high signals correspond with biological effects. High signals are also caused by experimental noise and common binding characteristics. Peak calling relies on correctly chosen parameters to filter noise from signal and is likely to classify weak binding motifs and co-factors as noise. Additionally, motif discovery is accomplished with a tool that is detached from the original data source which introduces another source of error.
Methods and results
We propose a new method to discover motifs directly from the integration profile of k-mers based on mapped reads. For each k-mer we cumulate the reads around all occurrences of the k-mer in the genome. The resulting profiles of protein binding k-mers show a characteristic shape. k-mers are filtered and scored according to their profile and then combined to sequence logos. This method offers the possibility to directly include experiment specific background data to remove noise.
Discussion
Our software NoPeak is able to find known motifs, has the ability to find weak binding motifs in ChIP-Seq data and is less susceptible to wrongly chosen parameters.

Using Nanopore Sequencing for in vivo Cell Type-Specific RNA Expression Profiling

Till Baar [(1)], Michael Ignarski [(2)], Sebastian Dümcke [(1)], Isabell Helmuth [(3)], Jasmin Hertler [(3)], Mark Helm [(3)], Roman-Ulrich Müller [(2)], Achim Tresch [(1)]
(1) Institute of Medical Statistics and Computational Biology IMSB, University of Cologne, Germany
(2) Dept. II of Internal Medicine and Center for Molecular Medicine Cologne, Germany

(3) Institute of Pharmacy and Biochemistry, Johannes Gutenberg-University Mainz, Germany

Transcriptome analysis has seen dramatic technological advances, starting with microarrays, followed by high throughput short read sequencing, and the development of single molecule sequencing technologies, such as Oxford Nanopore or Pacific Biosciences. Oxford Nanopore Technologies has presented a microfluidic device allowing the direct (i.e., without cDNA conversion), full-length sequencing of RNA. The transcript of interest is threaded through a protein pore, with the nucleotides inside the pore influencing an ion current flowing through it due to a voltage difference. This ion current is measured and analyzed to identify the nucleotide sequence.
Using Oxford Nanopore technology, we aim to generate in vivo, cell type-specific RNA expression profiles. We achieve this by combining direct sequencing with metabolic labelling of RNA. The latter uses chemically modified nucleotide analogues (5-ethynyl-uridine, 5-EU) to label nascent RNA in a cell type-specific and inducible manner. Our approach avoids several drawbacks of previous methods resulting from RNA purification, reverses transcription and amplification, and thus leverages the quantification of cell-type specific RNA expression. Here, we demonstrate the feasibility of the approach. We have constructed RNA in which uridine has been replaced by 5-EU in 0%, 1%, 10%, and 100% of all occurrences. Using segmentation algorithms, the ion current signal is cut into consecutive chunks of individual data points, which serve the identification of the k-mers (k=5 or 6) which reside in the pore. The sequence of k-mers is analyzed with machine learning strategies such as support vector machines (SVMs), in order to discriminate labeled from unlabeled k-mers. Then, we combine the evidence of single k-mers into a decision on whether or not the whole read is labeled. We benchmark the sensitivity and specificity of our call, and show that we can generate reliable expression profiles of the labeled (i.e., cell-type-specific) RNAs.

BRAKER2: Incorporating Protein Homology Information into Gene Prediction with Genemark-EP and Augustus

Katharina Hoff [(1)], Alexandre Lomsadze [(2)], Mario Stanke [(1)], Mark Borodovsky [(2)]
(1) University of Greifswald, Germany
(2) Georgia Institute of Technology, United States

The rapidly growing number of sequenced genomes requires fully automated methods for accurate gene structure annotation. With this goal in mind, we had developed BRAKER1, a combination of GeneMark-ET and AUGUSTUS, that uses genomic and RNA-Seq data to automatically generate full gene structure annotation in novel genomes. BRAKER2 is an extension of BRAKER1 which allows for the training and prediction on the basis of protein homology information, either alone or in combination with RNA-Seq evidence. If input RNA-Seq data, BRAKER2 runs self-training GeneMark-ET supported by the RNA-Seq alignments, trains AUGUSTUS on the basis of GeneMark-ET predictions, maps protein sequences of related species to the genome in question and incorporates both the RNA-Seq and protein information into gene prediction with AUGUSTUS. In absence of RNA-Seq data, BRAKER2 executes self-training GeneMark-EP supported by protein data, trains AUGUSTUS on the basis of GeneMark-EP predictions and predicts genes with protein homology information with AUGUSTUS. If the proteome of a very closely related species is available, BRAKER2 can even be more accurate without RNA-Seq than BRAKER1 with RNA-Seq.

Network analysis of DNA methylation of visceral adipose tissue comparing obese individuals with and without type 2 diabetes

Afsaneh Mohammadnejad [(1)], Weilong Li [(1)], Jesper Lund [(1)], Jan Baumbach [(2)], Qihua Tan [(1)]
(1) University of Southern Denmark, Denmark
(2) Technical University of Munich, Germany

Background: Type 2 diabetes (T2D) is a chronic condition in which the cell shows resistance in response to insulin. However, the underlying molecular mechanism of obesity and T2D is not fully understood. In this study, we performed a weighted gene correlation network analysis (WGCNA) in the methylome of human visceral adipose tissue (VAT) of obese individuals, to identify important genes from highly correlated network modules and functional pathways associated with T2D.
Methods and results: First, we summarized more than 485,000 methylation sites to gene level. Next, the WGCNA was conducted on gene levels from 18 (8 T2D and 10 nonT2D) female samples. Based on the methylation patterns we clustered the genes into modules and then correlated them to T2D. To find the biological function of genes and pathways, gene ontology (GO) software was applied and gene-gene interaction network of top genes in interesting modules was visualized in VisANT software. We identified two significant network modules, each consisting of 152 and 59 genes which are positively correlated with T2D. Furthermore, several hub genes such as CHCHD3, CFD, PABPC5, BMP6, EIF5, and EPS15L1 were detected, in which some are related to obesity, T2D, and the metabolic syndrome. Additionally, mitotic checkpoint, anaphase-promoting complex and some general metabolic pathways were significantly enriched from GO findings.
Discussion: Our network-based analysis revealed important genes within distinct modules that might be related to T2D in obese individuals. These findings suggest that there are methylation differences and significant contribution of epigenetic factors which might help to elucidate the underlying mechanisms of T2D.

Quantitative Prediction of Transcription Activator-Like Effectors

Annett Erkes [(1)], Maik Reschke [(2)], Stefanie Mücke [(2)], Jens Boch [(2)], Jan Grau [(3)]
(1) Martin Luther University Halle–Wittenberg, Germany
(2) Department of Plant Biotechnology, Leibniz Universität Hannover, Germany
(3) Institute of Computer Science, Martin Luther University Halle-Wittenberg, Germany

In many parts of the world, the cultivation of rice is important for ensuring nutrition of the population. Plant-pathogenic Xanthomonas bacteria cause diseases on various crop plants including rice, where Xanthomonas infections can lead to a harvest loss of up to 50%. Thus, it is important to investigate
the pathogenicity of these bacteria in more detail to find ways to protect the rice plants or to breed resistant ones.
Xanthomonas bacteria express proteins called transcription activator-like effectors (TALEs) that bind to the promoter of plant genes and activate their transcription. The binding domain of TALEs consists of tandem repeats of approximately 34 amino acids. Each repeat contains two hypervariable amino acids at positions 12 and 13, which are called repeat variable di-residue (RVD). Each RVD recognizes one nucleotide of its target DNA and the consecutive

array of RVDs determines TALE target specificity. The target sequence usually begins with nucleotide T at a zero-th, cryptic repeat.
Here, we present a new method to improve the prediction of TAL effector binding sites, which models more complex dependencies than previously available tools. Increasing model complexity is possible due to the availability of much larger, quantitative training datasets collected recently. We model the binding of a TALE to a putative target box using independent terms for i) the dependency of the zero-th
nucleotide on the first RVD, ii) the binding between the first RVD and the first nucleotide of the target sequence and iii) the binding of the remaining RVDs to the respective nucleotide of the target sequence, where the latter may also depend on the previous RVD. In addition, we weight these independent terms based on the position within the target sequence.
We benchmark this novel method against existing approaches on a variety of public and in-house RNAseq data sets measured in rice after Xanthomonas infection and find a generally improved prediction performance on the level of target genes as well as TALEs with at least one true positive target prediction.
Prediction performance and runtime allow us to use this new prediction tool for genome-wide binding site prediction of different TALEs from Xanthomonas oryzae pv. oryzicola (Xoc) and Xanthomonas oryzae pv. oryzae (Xoo) in rice. To this end, we scan the rice genome for putative binding sites and checked the RNA-seq data for differential expression in surrounding regions. While in some cases, differential expression is observed for rice genes that are known TALE targets, we also observed clear patterns of differential expression in regions without annotated genes. The latter could either be due to missing gene annotations in the rice genome or due to transcriptional activation by TALEs at genomic sites with imperfect transcriptional start sites.

CircRNA profiling in human cancer cells under hypoxic stress

Antonella Di Liddo [(1)], Camila Freitas Stahl [(2)], Sandra Fischer [(3)], Stefanie Ebersberger [(4)], Stefanie Dimmeler [(2)], Julia E. Weigand [(3)], Michaela Müller-McNicoll [(2)], Kathi Zarnack [(1)]
(1) Buchmann Institute for Molecular Life Sciences (BMLS), Goethe University Frankfurt, Germany
(2) Goethe University Frankfurt, Germany
(3) TU Darmstadt, Germany
(4) IMB Mainz, Germany

Hypoxia occurs when tissues are deprived of adequate oxygen supply and is associated with several diseases, including cancer. The response to hypoxia is extensively regulated at the transcriptional level and posttranscriptional level, including alternative splicing. Circular RNAs (circRNAs) are a novel class of long non-coding RNAs that are produced through an atypical splicing reaction referred to as back-splicing. So far, little is known about the biogenesis and function of circRNAs in hypoxia.
In order to characterise the circRNA profile of human cancer cells and their response to hypoxia, we performed total RNA sequencing (RNA-Seq) of cervical (HeLa), breast (MCF7) and lung (A549) cancer cells, subjected to hypoxic and normoxic conditions. We established a computational pipeline to reliably detect circRNAs by integrating two available tools, find_circ and CIRCexplorer, combined with custom approaches for quantification and statistical analysis. Using this consolidated pipeline, we identified more than 12,000 circRNAs,

of which several candidates were validated by RT-PCR and RNase R treatment. Circular-to-linear ratio analysis revealed that a number of circRNAs are more abundant than their linear counterpart. Overall, we observed a weak correlation between circRNA and host gene expression, indicating that circRNA biogenesis depends on independent regulatory features. Comparing normoxic to hypoxic conditions, we found that 64 circRNAs significantly change in abundance, with the vast majority being regulated in a cell type-specific manner. The notable exception was cZNF292 which was consistently upregulated in two out of three cell lines and previously shown to modulate the hypoxia response in endothelial cells. We confirmed the regulation of cZNF292 and six additional circRNAs using qPCR. In summary, we present a comparative profiling of circRNAs in different cancer cell lines and their response to hypoxic stress, which promises novel insights into the biogenesis and function of circRNAs in the future.

Finding Syntenic Regions in Multiple Unannotated, Unaligned Genomes

Matthis Ebel [(1)], Ingo Bulla [(2)], Mario Stanke [(1)]
(1) University of Greifswald, Institute of Mathematics and Computer Science, Germany
(2) Université Perpignan Via Domitia, IHPE UMR 5244, CNRS, France

We present a new approach for finding tuples of homologous regions in multiple genomes. Our aim is to improve de novo comparative genome annotation of clades. We target clades of many related genomes that have often undergone genome rearrangements since the most recent common ancestor and have low sequence conservation in non-coding regions. They are, however, evolutionary close enough to be locally alignable in principle.
Other approaches for the task of finding longer homologous regions in multiple genomes either require annotated input genomes or build on a multiple genome alignment. In contrast, our approach works with unannotated and unaligned genomes. It is based on the idea of geometric hashing, a technique to detect recurring patterns in data that may have undergone affine transformations. We apply geometric hashing on a set of k-mer occurrences. These are short sequences that have exact matches in multiple genomes. They can be directly queried from a suitable genome graph data structure. A k-mer occurence defines a potential relative offset between homologous regions of the genomes. In order to gain statistical power we seek to identify many k-mer occurences with similar relative offsets. With geometric hashing, the k-mer occurences are transformed to points in an (s-1)-dimensional space where s is the number of input species. Such a point denotes relative distances of the underlying k-mer sequence positions with respect to a reference genome. These points are then quantized to identify sets of multiple k-mers of similar relative distances. Finally, candidates for homologous regions over multiple genomes can be extracted from such sets.
In subsequent steps we plan to extend these homologous region tuples to larger syntenic regions. Those are homologous regions that span multiple homologous genes in the same order. These shall then aid the de novo comparative genome annotation of clades from which we expect to get a significant improvement of the resulting annotations.

3DPatch: tools for fast 3D structure visualization with residue conservation

David Jakubec [(1)], Rob Finn [(2)], Jiri Vondrasek [(1)]
(1) Institute of Organic Chemistry and Biochemistry of the CAS, Czechia
(2) EMBL-EBI, United Kingdom

Background: Amino acid residues manifesting high levels of conservation are often indicative of functionally significant regions of protein structures. For example, residues critical for protein folding, hydrophobic core stabilization, intermolecular recognition, or enzymatic activity often manifest lower mutation rates compared to the rest of the protein. Quantitative assessment of residue conservation typically involves querying a sequence against a database, finding similar sequences, aligning them to bring equivalent positions into register, and applying an information theory-based measure to individual columns in the multiple sequence alignment. Understanding how the sequence conservation profile relates in three dimensions (3D) requires its projection onto a protein structure, which can be a time-consuming process.
Methods and results: We developed 3DPatch, a client-side web application that simplifies the task of calculating protein sequence residue-level information content, homologous 3D structure identification, and conservation level-based mark-up. 3DPatch utilizes the power of profile hidden Markov models and speed of HMMER3.1 to provide accurate results in a matter of seconds. It was developed with easy integration into other peoples' websites in mind and supports most modern web browsers. 3DPatch is fully open-source and is freely available at http://www.skylign.org/3DPatch/. The 3DPatch web application is complemented by a set of command line tools (https://github.com/davidjakubec/3DPatch-tools) which present a scalable approach to performing residue-level conservation annotation for large sets of protein sequences and 3D structures.

Population structure and allele frequency clines in wild and cultivated beets (Beta vulgaris)

Jamie McCann [(1)], Felix Wascher [(1)], Nancy Stralis-Pavese [(1)], Britta Schulz [(2)], Juliane Dohm [(1)], Heinz Himmelbauer [(1)]
(1) University of Natural Resources and Life Sciences (BOKU), Austria
(2) KWS SAAT SE, Germany

Beta vulgaris (beet) is an economically important crop plant, whose cultivation is responsible for a significant proportion of the world's sugar production. Herein, genome resequencing data from multiple sugar beet and sea beet (the wild progenitor of beet cultivars) accessions, was analyzed. First, an alignment-free approach (MASH), using kmer sketches to obtain an unbiased estimate of the Jaccard index as a proxy for genetic distance, was applied to the sequencing data. The results provide evidence for two subpopulations of sea beet, one restricted to the Mediterranean and the other on the Atlantic Coast. A second approach using read mapping to the sugar beet reference (bowtie2) and SNP calling was also applied to the data. A program for estimating effective migration surfaces (EEMS), taking squared Euclidean distance of allele frequencies and geographic coordinates between pairs of georeferenced sea beets into account, was then used to determine potential barriers to gene flow. The results indicate a strong continental migration barrier separating the wild beet accessions, which corroborates the findings of the alignment-free approach.

The inference of information transfer in biological systems

Javier Geijo [(1)], Thomas Rattei [(1)]
(1) Universität Wien, Austria

Despite the crucial importance of microbial communities in the environment and for complex species, including us humans, their interaction networks and the functions of their members are mainly unknown. These ecosystems engage in complex trophic webs based on interspecies interactions, being each specie influencing others in bigger or lower degree along time.
Considering microbial communities a multi component system with different species abundances along time, important information on its structure can be obtained by measuring to which extent the individual components contribute to information production and at what rate they exchange information among each other. Transfer entropy, an information theoretic measure, is able to quantify information exchange between systems and its direction. It can be applied to the study of microbial dynamics in a system when several time series measurements (i.e. microbial abundances) are available.
The generation of genome-scale metabolic reconstructions of whole bacterial species and the subsequent curation of genome-scale metabolic reconstructions of whole bacterial communities is possible thanks to metabolic modelling approach. This approach allows the in-silico simulation of bacterial communities and it evolution across the time, being possible to offer enough number of time series measurements in order to infer the Transfer Entropy, being these in-silico simulations an ideal case study to test information theoretic measurements to complex systems.

By simulating an in silico human gut microbiota, entropy transfer has been inferred along the time. In addition, levels of potential, connectedness, and resilience were measured revealing a process of change common in the complex adaptive systems. In conclusion, metabolic modelling and its subsequent in silico simulations could be an ideal playground to test the viability of information transfer calculation in biological systems before its application on real data.

Extensions to Peptide Spectrum Match Validation by Semi-Supervised Machine Learning Methods

Georg J. Pirklbauer [(1)], Stephan M. Winkler [(1)], Karl Mechtler [(2)], Viktoria Dorfer [(1)]
(1) University of Applied Sciences Upper Austria, Bioinformatics Research Group, Softwarepark 11, 4232 Hagenberg, Austria, Austria
(2) Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Campus-Vienna-Biocenter 1, 1030 Vienna, Austria

A substantial part of proteomics mass spectrometry experiments aim at the identification of peptide sequences. This is often achieved by sequence database searching, resulting in peptide-spectrum matches (PSMs). Validation of PSMs is a crucial topic in the community.

Attributing confidence to PSMs allows for the retention of only statistically relevant identifications. Searching a target and a decoy database emerged as a practical way of estimating confidence and is universally accepted in the literature [1].
Based on the target-decoy approach, Käll et al. described a statistical framework that allows for the imputation of statistically sound confidence scores [2]. Based on this scoring they developed Percolator, an algorithm which allows boosting the number of confidently identified peptides at an arbitrary false positive rate cutoff [3]. The algorithm relies on a support vector machine and is widely accepted as a standard post-processing procedure in proteomics mass spectrometry experiments.
Since the development of Percolator, alternatives to the support vector machine have been developed and brought to maturation. We believe that an increase in the number of PSMs can be achieved by combining ideas of the Percolator algorithm and new machine learning techniques. We utilised random forests [4] in an iterative approach similar to the Percolator algorithm. Compared to the standard target-decoy approach we were able to increase the number of confidently identified PSMs at 1% FDR by 18% on a standard HeLa sample.
[1] J. E. Elias and S. P. Gygi, “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry,” Nat. Methods, vol. 4, p. 207, Feb. 2007.
[2] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble, “Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases,” J. Proteome Res., vol. 7, no. 1, pp. 29–34, Jan. 2008.
[3] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss, “Semi-supervised learning for peptide identification from shotgun proteomics datasets,” Nat. Methods, vol. 4, no. 11, pp. 923–925, Nov. 2007.
[4] T. K. Ho, “Random decision forests,” in Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, vol. 1, pp. 278–282 vol.1.

Small-scale phasing of somatic variants for neoantigen discovery using Microphaser

Jan Rouven Forster [(1)], Johannes Köster [(2)]
(1) Deutsches Konsortium für translationale Krebsforschung, Standort Essen/Düsseldorf, Germany
(2) Algorithms for reproducible bioinformatics, Institute of Human Genetics, University of Duisburg-Essen, Germany

One of the biggest challenges in cancer immunotherapy lies in the design of functional neoantigens.
A crucial step in neoantigen prediction is the generation of a cancer-specific neo-peptidome. Hereby, current approaches mostly rely on incorporating somatic variants into the human reference. In the light of precision oncology, it seems promising to further include patient-specific germline variants as well as peptide phasing to create a more specific and accurate search space for neoepitope candidates. Therefore, we introduce Microphaser, a tool for fast (linear-time) small-scale phasing of cancer peptides.
Microphaser slides over the reading frames of coding regions in windows of typical epitope length (9 codons). We consider all variants in a window and compute sequence and frequencies of supported haplotypes over all reads that enclose this window.
Haplotypes that contain at least one somatic variant are taken as candidates for neoantigen

prediction, all other haplotypes are used as background for filtering self-similarity in neopeptides.
On a dataset of 17 malignant pleural mesothelioma patients, we could identify several neoantigen candidates which were not found using standard prediction pipelines and showed an improved HLA binding affinity compared towards their most similar wildtype peptides.
Furthermore, we increased the patient-specific healthy peptidome by including germline variants, leading to a higher chance of filtering highly self-similar false positive neoantigen candidates.
Therefore, it can be reasoned that incorporating germline as well as somatic variants and small-scale phasing improves the accuracy of both healthy and cancer peptidomes and could improve the discovery of possible neoantigens for personalized immunotherapy.
(Code at https://github.com/koesterlab/microphaser)

Modelling Dependencies Between Histone Modifications Using Deep Learning

Christian Thomae Viegas [(1)], Andreas Dominik [(2)], Andreas Gogol-Döring [(2)]
(1) German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany
(2) Technische Hochschule Mittelhessen, Germany

Background
Histone modifications are one of the epigenetic mechanisms which may regulate gene expression but are not yet fully understood.
We used deep neural networks to model connections between different histone modifications, as they have extensive representation and approximation capabilities.
The learned implicit model enables us, for example, to predict missing data about histone modifications for one cell type based on interdependencies between histone modifications as they could be inferred from data of other cell types.
Methods and results
Using sample data from the ENCODE project with cell types GM12878 for training, H1-hesc and HeLa S3 for validation, it is possible to create a model that learns the interdependencies between different histone modifications.
Our model electively utilizes not only the interdependencies but also the spatial context around the prediction position.
For most histone modifications the model achieves a prediction quality that is in the range of biological replicates although there is little similarity between the used cell types.
Validation on data of cell types H1-hesc and HeLa S3 results in a mean sensitivity of 65% or 73% resp., a specificity of 95% or 82% resp., and an average AUC of 0.88+-0.11 or 0.85+-0.09 resp.
Discussion
While the resulting model reaches an reasonable prediction quality, a qualitative analysis of the predictions and a quantitative comparison with existing approaches is left for future research.

An organism-specific hydrophobicity scale based on reference datasets

Martin Ortner [(1)], Mark Teese [(1)], Dieter Langosch [(1)]
(1) Technical University of Munich, Germany

Hydrophobicity scales are a key feature to predict transmembrane (TM) regions of proteins. However, one major drawback of most published experimentally determined scales is their dependence on one specific organism.
In silico modelling of α-helical TM protein segment evolution was performed using human codon usage as amino acid (aa) propensities to mutate residues. Mutational positions were randomly distributed across the full TM segment. The lipophilicity of the generated sequence was obtained using different published hydrophobicity scales. A strict hydrophobicity cut-off was applied to ensure membrane insertion of the generated sequence. If the hydrophobicity of a resulting segment was not sufficient to maintain membrane insertion, it was rejected, and the previous sequence was used as template for the next mutational event. The performances of published hydrophobicity scales were tested using our model at varying cut-offs and comparing corresponding aa propensities to a reference dataset containing human bitopic membrane proteins. We calculated the mean difference between the aa propensities across all 20 aa and determined the optimal hydrophobicity cut-off value that showed the closest difference between the generated sequences and our reference. We could show that hydrophobicity scales have to be asymmetrical for good performance. Furthermore, we can show a concept to generate novel organism-specific hydrophobicity scales based on a reference dataset and the organism's codon usage.
We propose that published hydrophobicity scales alone cannot depict the hydrophobicity of amino acids across all species. Our novel method to generate organism-specific hydrophobicity values did not only show the closest match to the reference dataset in our model but is also easy to apply in all kinds of algorithms. In future versions, changing hydrophobicity values based on the position of a residue according to its environment are possible.

Complexity and Gene Enrichment in Vertebrate Genomes

Jessica Christin Piontke [(1)], Bernhard Haubold [(1)]
(1) Max Planck Institute for Evolutionary Biology, Germany

Background:
Unique regions of vertebrate genomes are often associated with biological function.
We search for such regions by computing the local complexity of DNA sequences across whole genomes sampled from the full diversity of the vertebrate clade.
Methods and Results:
Our metric for sequence complexity is the match complexity, C_m, implemented in the program macle. C_m = 0 for repeated regions, and C_m =~ 1 for random DNA. For each genome analyzed, we use macle to compute C_m in 10 kb sliding windows. Averaging these windows across transcription start sites (TSS) shows that C_m peaks near the TSS and falls off asymmetrically in such a way that it is lower on the 5' side than on the 3' side, where the gene is located. The relative height of this peak varies between less than 1% in Gallus gallus and 6% in Bos taurus.

Moreover, we find that regions with C_m =~ 1 are significantly enriched for genes. The extent of gene enrichment varies in mammals between 1.6-fold in Homo sapiens and 2.6-fold in Rattus norvegicus and decreases from mammals over birds and amphibians to fish.
Discussion:
We have shown that a simple measure of sequence complexity, C_m, is strongly associated with biological function. The next step is to investigate whether any functional categories are enriched in the genes of particular complexity.

An improved scoring scheme for local sequence-structure alignments to identify regulatory RNAs

Teresa Müller [(1)], Milad Miladi [(1)], Sebastian Will [(2)], Ivo Hofacker [(2)], Rolf Backofen [(1)]
(1) Bioinformatics, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany, Germany
(2) Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria, Austria

Background:
Alignments of molecular sequences allow us to infer their biological functions. For comparing RNAs, high identity sequence alignment can provide accurate results. However, structure can be more conserved than sequence as functionality of a regulatory RNA is generally associated with its structure. Current sequence-structure alignment tools compute acceptable alignments in global mode. In contrast, local alignment has remained a challenge when searching for structured regulatory motifs that are embedded in their genomic sequence contexts.
Methods and Results:
In this work, we explore the scoring function of LocARNA (Will et al. 2007), as a state-of-the-art sequence-structure alignment method. The main goal is to identify the contribution of RNA structure on the similarity score for pairwise alignment and how it can be included more effectively. Using an artificial data set in which we control sequence length and nucleotide frequencies, we show that the structure information of sequence-structure alignments creates a positive scoring bias on average. This bias leads to longer local alignments for shuffled functional homologs. We propose an extension to the local scoring scheme that compensates the positive structure scoring bias. Initial benchmarking experiments show improved results on BRAliBase alignments (Wilm et al. 2006) that were extended by their shuffled genomic context to simulate a local setting.
Discussion:
Our proposed scoring scheme for local alignment not only improves LocARNA results but should be applicable to methods using a similar objective function.

Will, S. et al., 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS computational biology, 3(4), p.e65.
Wilm, A., Mainz, I. & Steger, G., 2006. 10.1186/1748-7188-1-19. Algorithms Mol Biol, 1(1), p.19. Available at: http://almob.biomedcentral.com/articles/10.1186/1748-7188-1-19.

RiboSNitches in complex cardiovascular diseases

Leonie Martens (1), Frank Rühle (1), Monika Stoll (1)
(1) Department of Genetic Epidemiology, Institute of Human Genetics, University of Münster, Germany, Germany

Leonie Martens[1], Frank Rühle[1], Monika Stoll[1,2]
[1]Department of Genetic Epidemiology, Institute of Human Genetics, University of Münster, Germany
[2]Department of Biochemistry, Genetic Epidemiology and Statistical Genetics, CARIM School for Cardiovascular Diseases, Maastricht Center for Systems Biology (MaCSBio), Maastricht University, The Netherlands

Background
Genome wide association studies contribute to the understanding of the genetics of complex diseases, with a significant proportion of the association signals residing in non-coding regions. Long non-coding RNAs (lncRNAs) have been shown to affect gene regulation in multiple ways and play an essential role in disease. We set out to identify novel riboSNitches, a genetic variant altering RNA structure, involved in complex cardiovascular disease.
Methods and results
We performed targeted sequencing in 96 DNA samples from cardiomyopathy patients comprising a genomic region of 540 kb including lncRNAs previously associated with heart disease. We identified 4636 genomic variants possibly affecting their secondary structure. Subsequently, we investigated relevant RiboSNitches for disease association in a population-based study and filtered them for expression of the corresponding lncRNA in the human heart using publicly available RNASeq data. The impact of a variant on the resulting secondary structure was estimated by a combination of established prediction algorithms. We compared the minimum free energy structure changes and developed a method to measure the impact of the genomic variant on the structural Boltzman ensemble. The most promising candidates were identified in lncRNAs Bigheart, Carmn, CDKN2B-AS1, H19, HCG22 and MLIP-AS1, which are now subject to experimental validation by SHAPE-Seq chemical probing.
Discussion
Preliminary results suggest that lncRNAs, previously shown to play an important role in cardiovascular diseases, are affected by riboSNitches. In silico prediction methods suggest a significant impact of these variants on the secondary structures and their contribution to disease etiology.

EXOMEDA – The new text mining approach for organism-related information in BRENDA

Sandra Placzek (1), Dietrich Ober (2), Dietmar Schomburg (1)
(1) TU Braunschweig, Germany
(2) Christian-Albrechts-Universität zu Kiel, Germany

The continuously increasing number of scientific publications in the area of the Life Sciences necessitates the intensified use of text mining procedures to extract the most important information. For more than ten years text mining methods have been used by the BRENDA team to automatically extract a variety of information on enzymes, including the source organisms and tissues, as well as localizations and kinetics from the scientific primary

literature (www.brenda-enzymes.org).
Since three years BRENDA – now an ELIXIR Core Data Resource (https://www.elixir-europe.org/platforms/data/core-data-resources) - is part of de.NBI, the German network for bioinformatics infrastructure (www.denbi.de). During this time a new text mining approach has been developed to automatically extract organism-related metadata of the categories “organism”, “metabolite”, “enzyme”, “metabolic pathway”, “habitat”, “region”, “human disease”, “plant disease”, and “plant trait” from all organism-related PubMed titles and abstracts. Weighted co-occurrence analyses connect these features.
In order to provide the user with a quick overview of the information variety of organism metadata, taxonomic word maps have been designed based on BRENDA’s word maps for enzymes. These new word maps allow the user to recognize an organism’s scientific context without having to read a multitude of scientific publications or data tables. A map of the regions occuring together with an organism name in the scientific literature allows an overview of an organism’s geographic scientific relevance.
A summary page for organisms including the manually annotated data, the obtained text mining data, and the new word maps provides a quick overview of a desired organism.

Phylogenetic terraces and an efficient tree space exploration

Olga Chernomor [(1)], Lukasz Reszczynski [(1)], Arndt von Haeseler [(2)]
(1) University of Vienna, CIBIV, Austria
(2) University of Vienna, Medical University of Vienna, CIBIV, Austria

In phylogenomics, the analysis of concatenated gene alignments, the so-called supermatrix, is commonly accompanied by the assumption of partition models. Under such models, each gene, or more generally partition, is allowed to evolve under its own evolutionary model. Though partition models provide a more comprehensive analysis of supermatrices, missing data may hamper the tree search algorithms due to the existence of phylogenetic terraces - collections of trees with identical score (maximum likelihood or parsimony score).
For sparse supermatrices with a lot of missing data, the number of terraces and the number of trees on the terraces can be very large. If terraces are not taken into account, a lot of computation time might be unnecessarily spent to evaluate many trees that in fact have identical score. Thus, exploration of tree-space is inefficient and due to limitations of numerical accuracies, the trees on a terrace will have slightly different scores, another unwanted effect.
In our previous work, we provided an efficient way of saving computational time by identifying consecutive trees that lie on the same terrace, and generalized the concept to partial terraces, which occur more frequently than the original ‘‘full’’ terrace and provide additional timesaving possibilities.
In this poster, we provide further insights into combinatorial properties of phylogenetic tree search in the presence of missing data and terraces. Among others, we discuss how difficult it is to leave the terrace under random nearest neighbor interchange and whether it is useful to consider more than one tree from the terrace during the search. Such insights are valuable for the improvement of tree space exploration for contemporary phylogenomic alignments.
O. Chernomor, B.Q. Minh, and A. von Haeseler (2015) Consequences of Common Topological Rearrangements for Partition Trees in Phylogenomic Inference. Journal of Computational Biology, Dec; 22(12):1129-42. (DOI:10.1089/cmb.2015.0146)

O. Chernomor, A. von Haeseler, and B.Q. Minh. (2016) Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst Biol. (DOI:10.1093/sysbio/syw037)

Promising Serological Biomarkers to Diagnose Giant Cell Arteritis in a Large Cohort of Treatment-Naïve Patients

Julia Feichtinger [(1)], Gerhard Thallinger [(1)], Snezna Sodin-Semrl [(2)]
(1) Graz University of Technology, Austria
(2) University Medical Centre Ljubljana, Slovenia

Background
Giant cell arteritis (GCA) is the most common form of vasculitis in adults leading to inflammation of the blood vessels. GCA etiology is unknown and even diagnosis remains difficult, as it relies on invasive temporal artery biopsies or imaging modalities (vascular ultrasound). Therefore, we are currently working on determining panels of suitable GCA biomarkers for clinical use with the goal of enabling personalized disease management.
Methods and Results
To screen a large number of biomarker candidates, we have used a combination of methods, including systematic review and meta-analysis as well as statistical evaluation, clustering and principal component analysis of clinical data and analyte measurements (Luminex, ELISA and nephelometry data) in the largest cohort of treatment-naïve patients (n=97) to date. Thereby, we have identified promising GCA biomarkers including, among others, a number of acute phase parameters and interleukins. Our results show that a number of biomarkers are not only promising candidates for early diagnosis but also for predicting GCA complications, such as relapse and visual disturbances.
Discussion
Early diagnosis and treatment of GCA is crucial for preventing ischemic complications but, to date, there is a clear lack of serological markers for GCA. With medicine and pharma moving towards personalized medicine, the identification of biomarkers to enable just this is of great importance. The promising serological markers we identified could improve diagnosis and prognosis as well as enable a monitoring procedure for GCA patients, contributing substantially to personalized disease management.

PICA-to-go: A fast microbial phenotype investigation pipeline

Florian Piewald [(1)], Thomas Rattei [(1)]
(1) Computational Systems Biology (CUBE), University of Vienna, Austria

As the number of completely sequenced bacterial species grows, analyzing the phenotypes of these species becomes a bottleneck in science. Machine learning tools have been used in the past to cope with this problem. The PICA framework is an example of such a tool, using support vector machines. PICA, however, needs information about the COGs (clusters of orthologous groups) in each species (bin) to be trained/predicted. In previous work of our group, HMMER searching in the EggNOG database, was used for this purpose. The usage of this approach is limited to individuals with access to a computing cluster and takes a

considerably amount of time. I present a new approach (PICA-to-go) using the clustering suite of MMSeqs2 together with the PICA framework. A model (PEN_180) for penicillin resistance consisting of 180 bins of training data can be trained in less than ten minutes on an ordinary desktop machine.

Does activation of specific transposable elements play a role in the development of glioblastoma?

Konrad Grützmann [(1)], Falk Zakrzewski [(2)], Alexander Krüger [(3)], Daniela Aust [(2)], Evelin Schröck [(4)], Matthias Schlesner [(5)], David Jones [(6)], Julian Seufert [(7)], Liam Childs [(8)], Michael Weller [(9)], Jörg Felsberg [(10)], Bernhard Radlwimmer [(11)], Alfred Nordheim [(12)], Guido Reifenberger [(10)], Barbara Klink [(13)]
(1) National Center for Tumor Diseases (NCT) partner site Dresden, Germany;, Germany
(2) Institute for Pathology, University Hospital Carl Gustav Carus Dresden, Germany, Germany
(3) National Center for Tumor Diseases (NCT) partner site Dresden, Germany, Germany
(4) Institute for Clinical Genetics, University Hospital Carl Gustav Carus Dresden, Technische Universität Dresden, Germany, Germany
(5) Division of Theoretical Bioinformatics, Heidelberg; Bioinformatics and Omics Data Analytics, DKFZ, Heidelberg, Germany, Germany
(6) Hopp Children's Cancer Center at the NCT Heidelberg, Germany; Division of Pediatric Neurooncology, DKFZ, Heidelberg, Germany
(7) Bioinformatics and Omics Data Analytics, DKFZ, Heidelberg; Faculty of Biosciences, Heidelberg University, Germany, Germany
(8) Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany, Germany
(9) Department of Neurology, Clinical Neuroscience Center, University Hospital Zurich and University of Zurich, Switzerland, Germany
(10) Department of Neuropathology, Heinrich Heine University Düsseldorf, Germany, Germany
(11) Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany, Germany
(12) University of Tübingen, Institute of Medical Genetics and Applied Genomics, Tübingen, Germany, Germany
(13) National Center for Tumor Diseases Dresden; University Hospital Dresden, Technische Universität Dresden, Germany, Germany

Background
Glioblastoma is the most common primary brain tumor in adults. Glioblastomas usually return after treatment and affected patients have a very poor prognosis. No curative therapies are available so far. Tumor heterogeneity and tumor evolution pose an enduring challenge. Thus, better prognostic markers and novel therapeutic approaches are needed.

About 50% of the human DNA is comprised of transposable elements (TEs). Many TEs profoundly affect the genetic buildup e.g. via coding sequence disruption, epigenetic deregulation, and aberrant expression. For instance, mobility of Long Interspersed Nuclear Elements-1 (LINE-1) is typical for cancer, and can drive mutations during tumorigenesis.

Although little studied so far, TEs may contribute to gliolastoma evolution and could serve as prognostic markers.
Methods and Results
We performed ribosomal RNA-depleted sequencing of RNA from 53 glioblastoma patients from the German Glioma Network cohort and four normal brain controls. Tumor samples had significantly higher overall TE expression compared to normal brain samples. Furthermore, we found specific TE families and TE classes significantly deregulated in glioblastomas compared to normal brain controls. Among the strongest de-regulated families were putative DNA transposons (MER133A, MER131), endogenous retroviral sequence 1, and the uncharacterized UCON4. Moreover, the expression of specific long terminal repeat retrotransposons (LTR39, MER31B, LTR3B), LINEs (X7A_LINE, L1MCc, L1M3b) and DNA transposons (MER6, MER97b, Tigger15a) correlated with progression-free and overall survival.
Discussion
Our preliminary results suggest that the role of TE expression in tumor development should be further investigated. To approach the role of TEs in mutagenesis, we are currently developing a pipeline to reconstruct structural rearrangements between TEs and non-TE genomic loci.

Detection of Similar Strains in Metagenomic Assembly Bins from an Enrichment Culture

Heiko Schmidt [(1)], Anna Zappe [(1)], Ricardo J Eloy Alves [(2)], Christa Schleper [(2)]
(1) CIBIV, University of Vienna, Austria
(2) Archaea Ecology and Evolution, University of Vienna, Austria

Background
Metagenomic approaches are used to gain insights into what is out there in environmental samples like from soil, sea, piping, skin or gut.
Even to obtain genomic sequences of certain species of interest metagenome assemblies are often unavoidable. Typical reasons may be that species are not easily culturable or they depend on resources by other organisms in the sample.
In such cases, DNA extracted from such enrichment cultures is used for sequencing and the reads are assembled to contigs. Subsequently contigs need to be ‘binned’ to different putative species. However, proper assembly and binning is difficult if there are similar strains of species of interest are present.
Methods and results
We show a number of diagnostic approaches to check assembly bins. We use standard Bioinformatics tools for sequence comparison and visualizations to determine whether binned genomes contain mixtures of similar strains.
We show a biological example of a metagenome assembly to gain a Thaumarchaeote genome from sequencing reads from enrichment cultures. We were able to detect the presence of two very similar strains and finally sort them into two separate genomes.
Discussion
After binning contigs, a bin might contain very similar but still distinct strains. While sequencing of different samples/libraries from enrichment cultures strongly improves the assembly process and the binning of the contigs into OTUs, it increases the danger of introducing similar strains. Thus, we suggest performing some simple diagnostic tests that

help to detect such cases of mixtures and may help to sort out the closely related genome sequences.

Design and development of a comprehensive online resource for the biotechnological important yeast Komagataella phaffii

Nadine Tatto [(1)], Minsoka Valli [(1)], Alexandra Graf [(2)], Diethard Mattanovich [(3)]
(1) acib GmbH, Austria
(2) FH Campus Wien, Austria
(3) University of Natural Resources and Life Sciences, Austria

The annotation of the Komagataella phaffii CBS7435 genome went through a far-reaching re-annotation process, which was carried out
manually over great extent. The annotation process was backed substantially by new RNA-Seq data of this Komagataella strain.
This was a great opportunity to implement a new platform to fulfill the need of the scientific K. phaffii/Pichia pastoris community and to give access to
this new information. Therefore, in close collaboration with the annotating scientists and the potential user group, a website was developed,
based on a newly designed database, to grant this access. Additionally - to provide a more visualized view of the new data - a new genome
browser was installed and configured to meet the needs of the community.

Annotation of viral polyproteins

Jean Mainguy [(1)], Hans-Jörg Hellinger [(2)], Thomas Rattei [(2)]
(1) CUBE University of Vienna, Austria
(2) Cube, University of Vienna, Austria

Background:
Polyproteins are large proteins found in viral genomes. They contain several functional units, which are cleaved by host or viral proteases into biochemically active mature peptides. Polyprotein annotation across viral genomes in public database is mainly sparse and sometimes inconsistent. Computational prediction of mature peptides is very limited so far, but would be highly desirable for comparative genomics of viruses. Here we present the prototype of a method to automatically annotate polyproteins.
Methods and Results:
The core idea for the prediction is the propagation of cleavage sites from already annotated polyproteins to unannotated ones. Viral proteins are first extracted from the RefSeq database and clustered into homologous groups. The groups containing annotated and unannotated polyproteins are aligned. The quality of the alignment and the consistency of cleavage site annotations are evaluated automatically. Domain annotation boundaries are used to identify conflicts between cleavage site annotations and putative protein function. Finally, the cleavage sites are propagated from annotated to unannotated proteins, and the confidence score of the predictions are determined.

Discussion:
This method allows for the first time to annotate confidently the majority of the unannotated polyproteins of the RefSeq database. However, annotation errors can be propagated too as the approach relies solely on the already existing annotation. This risk can be controlled by the user by defining the cutoff for the confidence score, as well as manual assessment of conflicts between annotated cleavage sites and functional domain annotations.

Comparative genomics of the pathogenic bacterium Tannerella forsythia

Nikolaus F. Zwickl [(1)], Nancy Stralis-Pavese [(1)], Christina Schäffer [(2)], Heinz Himmelbauer [(1)], Juliane C. Dohm [(1)]
(1) University of Natural Resources and Life Sciences (BOKU), Department of Biotechnology, Vienna, Austria, Austria
(2) University of Natural Resources and Life Sciences (BOKU), Department of Nanobiotechnology, Vienna, Austria, Austria

Tannerella forsythia is a prokaryotic species associated with the polymicrobial disease periodontitis, the most common reason for tooth loss world-wide. Unraveling the species' contributions to pathogenesis as well as the underlying genetic basis has been hampered by the organism's fastidious growth requirements and slow growth, on one hand, and by human errors made in the context of adapting research on this organism to the genomics era. Here, we report an improved version of the genome assembly for the reference strain T. forsythia ATCC 43037 and compare the genome sequence to other T. forsythia strains as well as to its closest known relative, a periodontal health-associated species. We performed multiple whole genome alignments based on iterative detection of sequence blocks lacking internal rearrangements (progressiveMauve), a pan-genome analysis employing clustering of putative orthologs with OrthoMCL and COG algorithms based on shared sequence similarity and PFAM domains, and a codon usage analysis by means of the self-consistent normalized relative codon adaption index (scnRCA). We applied various custom scripts searching for common and diverged genomic architectures. The results enabled us to confirm previously reported virulence factors to be conserved throughout the species, underpinning these genes' or their products' suitability as anti-microbial targets. Comparing the species' genomic repertoire to the health-associated BU063 genome pointed to numerous genomic regions that might be involved in pathogenesis and therefore represent interesting topics for future research, which, in further consequence, could lead towards the development of novel therapeutic strategies.

Deep Learning for drug-induced liver injury prediction

Marco Chierici [(1)], Nicole Bussola [(1)], Margherita Francescatto [(1)], Giuseppe Jurman [(1)], Cesare Furlanello [(1)]
(1) Fondazione Bruno Kessler, Italy

Background
Drug-induced liver injury (DILI) is a major concern in drug development, as hepatotoxicity may

not be apparent at early stages but can lead to life threatening consequences. The ability to predict DILI from in vitro data would be a crucial advantage. In 2018, the Critical Assessment Massive Data Analysis (CAMDA) group proposed the CMap Drug Safety challenge focusing on DILI prediction.
Methods and results
The challenge data include Affymetrix GeneChip expression profiles for the two cancer cell lines MCF7 and PC3 treated with 276 drug compounds and empty vehicles, along with DILI binary labeling and a recommended train/test split for the development of predictive classification approaches. The microarray data was processed with the frozen Robust Multi-array Average (fRMA) method, treating train and test sets independently to avoid information leakage.

We developed Deep Learning architectures for DILI prediction on the challenge data and compared them to Random Forest and Multi-Layer Perceptron classifiers, with inputs either normalized and log-transformed gene expression of treated samples or expression log-fold change of treated samples vs vehicles.
All models were trained within the MAQC Data Analysis Plan, i.e. 10x5 cross-validation (CV) over the training set, stratified over classes, evaluating classification by Matthews correlation coefficient (MCC). The outcomes on both CV or test datasets of all machine learning models, including deep learning ones, were consistently low, with validation MCC below 0.2.
Discussion
Our results suggest that DILI prediction from the CMap challenge expression data alone is a difficult task.

Using continuous bag of words to interpret the hidden information of protein sequences in electron transport proteins

Yu-Yen Ou [(1)]
(1) Yuan Ze University, Taiwan

Deep learning is a subset of artificial intelligence and machine learning that uses multi-layer artificial neural networks to provide state-of-the-art performance in tasks such as object detection, speech recognition, and language translation. In bioinformatics, deep learning has also been used to transform biomedical data into valuable knowledge since the beginning of the 21st century.
In this study, we proposed a new method that uses continuous bag of words (CBOW) to interpret the hidden information of protein sequences in electron transport proteins from amino acid sequences, and the CBOW has been successfully applied in natural language processing research. The protein sequence is similar to an unknown language with 20 letters corresponding to 20 amino acids. An ordered chain of words may contain some hidden information about biological functions that contribute significantly to the classification of proteins.
The study was attempted directly from the amino acid sequence, we chose electron transfer protein, a group of proteins that transfer electrons during the metabolic function of cell function to do this. Therefore, studies of the identification and classification of electron transport proteins will help to improve understanding of the cell respiration system. Fasttext is used as a modeling tool that implements hashing techniques to achieve fast and memory

efficient mapping to perform this research. The sensitivity of the electron transport protein in the transport proteins is successfully identified as 60.53%, the specificity is 94.84%, the accuracy is 91.71%, and the MCC is 0.53. This method also improves the measurement metrics compared to previous related works. The proposed technique provides a web-based online tool for research purposes and also opens up promises for implementing natural language methods to solve problems in this field.

Peptidome – DB: A community-driven proteogenomics database to curate the peptidome

Sebastian Didusch [(1)], Thomas Rattei [(1)], Harald Marx [(1)]
(1) University of Vienna, Austria

Background

Proteogenomics leverages next generation sequencing (DNA, RNA) and mass spectrometry (MS)-based proteomics data to refine existing gene models and identify novel gene models, including open reading frames (ORF) encoding bioactive peptides, e.g small ORFs (smORFs) . Recent estimates indicate that hundreds of thousands of those ORFs may exist across the tree of life, that would notably expand the known peptidome. However, smORFs do not exhibit common gene structure signals, that are crucial to gene predictors. To improve efforts in structural genome annotation we build an integrated pipeline which results in a central repository for proteogenomic data and a complementary search database to Ensembl.

Methods and results

Peptidome-DB is an integrated Java pipeline comprising data preprocessing, peptide mapping, ORF clustering, proteogenomic classification, statistical validation and data storage. A web interface allows users to upload results from proteomic search engines and the respective search sequence database. The back end performs an exact string-matching (Aho-Corasick algorithm) on the peptides from the search engine to their sequence database entries. BLAST and Exonerate align entries against Ensembl databases from various molecular types (PEP, cDNA, ncRNA, DNA).Single-linkage clustering (SLC) groups ORFs from unannotated loci with exons from Ensembl PEP together. We compute an optimal SLC distance threshold from the intragenic and intergenic distance distributions from Ensembl. A decision tree classifies each cluster member into a proteogenomic event. Our preliminary results from a Medicago truncatula data set show high peptide mapping accuracy. We were able to map 283.562 out of 284.051 (99.82%) peptides to the latest Ensembl release.

Discussion

We developed a pipeline, Peptidome-DB, for the incorporation and curation of proteogenomic data. Peptidome-DB utilizes Ensembl genome assemblies to refine gene models and annotate the peptidome with protein-level evidence.