Poster abstracts of the 25th German Conference on Bioinformatics

Kay Nieselt[1,2], Nico Pfeifer[3,4], Andrei Lupas[1,2,5], and Oliver Kohlbacher[1,2,5]

[1] Center for Bioinformatics Tübingen, University of Tübingen, Tübingen, Germany
[2] Quantitative Biology Center, Tübingen, Germany
[3] Methods in Medical Informatics, University of Tübingen, Tübingen, Germany
[4] Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
[5] Max Planck Institute for Developmental Biology, Tübingen, Germany

The annual German Conference on Bioinformatics (GCB) has brought together renowned international scientists in bioinformatics, computational biology, biostatistics, biology and medicine since 1993. The 25th GCB was held in Tübingen from 18th to 21st of September 2017, making its second appearance at the vivid university city with its internationally renowned medical faculty as well as the faculty of science, and several top research institutes like the Max Planck Institute for Developmental Biology. This environment has created and attracted several internationally recognized research groups in bioinformatics, computational biology, and biomedical data science out of which many have contributed in organizing the GCB 2017.
Following are the abstracts of posters accepted for presentation at the conference, showing the wide range of different areas from bioinformatics infrastructure to theoretical approaches in single cell analyses.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

MPEG-G: The Emerging Standard for Genomic Data

Jan Voges and Jörn Ostermann
Institut für Informationsverarbeitung, Leibniz Universität Hannover
{voges,office}@tnt.uni-hannover.de

The development of next-generation sequencing (NGS) technologies enables the usage of genomic information as everyday practice in several fields. However, the growing volume of data generated becomes a serious obstacle for the advancement of related applications. To comprehend the volume of data that needs to be represented, stored and transmitted, current sequencing machines are capable of delivering over 18,000 whole human genomes a year, which accounts for almost 5 PB of data per year and system. Therefore, efficient storage and transmission of genomic data is becoming of uttermost importance. Besides compression, which is at the base of any efficient processing of genomic information, there are several other requirements that the current data formats do not fulfill [Joi16].

ISO/IEC JTC 1/SC 29/WG 11 — also known as Moving Picture Experts Group (MPEG) — has the mission to develop standards for coded representation and compression of digital audio, video and related data. In its 29 years of activity MPEG has developed many generations of audio and video compression standards such as MP3 and AVC (also sometimes referred to as H.264). ISO/TC 276 works on standardization in the field of biotechnology processes that include analytical methods (Working Group 3) and data processing and integration (Working Group 5). MPEG and ISO/TC 276/WG 5 (Data Processing and Integration) have combined their respective expertise and missions and are jointly working to develop a new and open standard for genomic information representation, called MPEG-G.

Data type	**Compression factor (approx.)**
Read identifiers	10
Quality values (lossless)	3.7
Quality values (quasi-lossless)	12.5 (with less than 3% F-score degradation)
Unaligned reads (constant & variable lengths)	25–58 (low to high coverage samples)
Aligned reads (constant length)	12
Aligned reads (variable lengths)	8

This standard will be offering higher levels of compression for all relevant data classes such as reads, quality scores/values, read identifiers, and alignment information [N+16]. The table shows the results of the initial technology performance assessment. The MPEG-G standard will furthermore provide new functionalities such as support for selective access, data protection mechanisms, conversion from/to the SAM/BAM file formats for backwards compatibility, and streaming of genomic data, enabling for example live streaming of data from a sequencing machine to a remote analysis center during sequencing.

The framework for the development of the open source standard MPEG-G is provided by ISO/IEC. Following the identification of requirements and the evaluation of technologies, the standardization process involves the selection and integration of the best performing technologies into a platform, called "General Model", for the evaluation and verification of performance and the validation of requirements fulfillment. This work started in January 2017 and is currently ongoing. The current Working Draft of the standard (ISO/IEC NP 23092) will evolve into a Committee Draft in October 2017 and a Final Draft International Standard in January 2019. Finally, a normative and informative specification (International Standard) in the form of text and reference software will be published. This specification will provide the foundation for interoperable genomic information processing applications enabling the use of genomic data on a large scale in fields such as personalized medicine, where the DNA of the patient will be sequenced and analyzed as part of a standard procedure.

[Joi16] Joint AhG on Genomic Information Compression And Storage. Requirements on Genomic Information Compression and Storage. Technical report, ISO/IEC JTC 1/SC 29/WG 11 (MPEG) and ISO/TC 276/WG 5, Document Number N16323/N97, Geneva (CH), 2016.

[N^{+}16] Ibrahim Numanagić et al. Comparison of high-throughput sequencing data compression tools. *Nature Methods*, 13(12):1005–1008, 2016.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Species-wide spectrum of resistance genes in Arabidopsis thaliana

Anna-Lena Van de Weyer, Felix Bemm, Freddy Monteiro, Oliver Furzer, Detlef Weigel
Max Planck Institute for Developmental Biology, Tuebingen
anna-lena.vd.weyer@tuebingen.mpg.de

Plant health is an essential component of crop yield. Plant researchers are thus driven to understand the molecular basis of plant immunity and resistance. Resistance genes are key players in a plant's fight against the tremendous diversity of pathogenic attackers. Nucleotide-binding and leucine-rich repeat (NLR) containing genes represent one of the most important resistance gene families in plants. They detect pathogenic effectors that try to interfere with cellular processes and induce resistance responses. As a result of an evolutionary arms race between plants and pathogens, NLRs have been shaped by repeated ancient and ongoing duplication events, with many NLR genes being found in complex clusters. High variability between strains has been inferred from comparisons of individual clusters for a small number of strains, but the true extent of species-wide NLR variation is unknown – even for the model plant *Arabidopsis thaliana*. Simple short read based re-sequencing approaches have largely failed to answer this question because of the excessive sequence and copy number variation between accessions. We have used instead NLR-sequence enrichment followed by long-read sequencing to assemble and annotate individual NLR'omes of a set of 65 *A. thaliana* accessions representing the global diversity of the species. Unexpectedly, a large fraction of genes was conserved and could be recovered by interrogating only a limited number of accessions. Some NLRs, however, are restricted to single accessions, or just a few accessions. Expression data from the 1001 Transcriptomes project was used to detect putative active NLRs that could be candidates for functional studies. We will discuss how rare NLRs, domain architecture differences, within-gene indels and SNPs contribute to NLR'ome variation. The structural description of the pan NLR'ome is a first step towards understanding the evolution of this important gene family in *A. thaliana*.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

In silico adaptive design of peptides with selective anticancer activity

Gisela Gabernet, Damian Gautschi, Alex T. Müller, Claudia S. Neuhaus, Jan A. Hiss and
Gisbert Schneider
Institute of Pharmaceutical Sciences, Swiss Federal Institute of Technology (ETH)
gisela.gabernet@pharma.ethz.ch

Membranolytic anticancer peptides (ACPs) are a promising strategy in the fight against cancer. Their receptor-independent mechanism of action is thought to hinder the development of resistances. Until now, hundreds of ACPs have been identified and collected in specialized databases. However, data on ACP selectivity towards non-cancer cells is not readily available and there is a lack of computational tools for the design of selective ACPs [GMHS16].

We aimed to develop a computational pipeline for predicting the activity of de novo generated ACPs and improving their selectivity towards non-cancer cells. We constructed a support vector machine classifier model that allowed to distinguish between anticancer active and inactive peptides. This model was used to analyse three different peptide libraries which we generated in silico [MGHS17]: (i) amphipathic peptides with varying hydrophobic arcs, (ii) peptides with a hydrophobic gradient along their sequence, and (iii) peptides with the amino acid composition of known alpha-helical ACPs. Selected peptides from each library predicted to be active or inactive, respectively, were synthesized and tested against two different cancer cell lines. 10 out of the 12 predictions turned out to be correct.

An evolutionary algorithm was then employed in order to improve the selectivity of the most active peptide [HSP+15]. After one run of peptide maturation, a 10-fold improvement in selectivity with regard to non-cancer cells, and a 15-fold improvement with regard to human blood erythrocytes was observed. The results of the present study provide proof-of-concept for the applicability of adaptive machine learning to ACP design.

References

[GMHS16] Gisela Gabernet, Alex T. Müller, Jan A. Hiss, and Gisbert Schneider. Membranolytic anticancer peptides. *Med. Chem. Commun.*, 7(12):2232–2245, 2016.

[HSP+15] Jan A. Hiss, Katharina Stutz, Gernot Posselt, Silja Weßler, and Gisbert Schneider. Attractors in Sequence Space: Peptide Morphing by Directed Simulated Evolution. *Molecular Informatics*, 34(11-12):709–714, nov 2015.

[MGHS17] Alex T. Müller, Gisela Gabernet, Jan A. Hiss, and Gisbert Schneider. modlAMP: Python for antimicrobial peptides. *Bioinformatics*, btx285, may 2017.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

ROMA: Ramachandran Oriented Mutational Analysis

Leonardo Alves Santos[1], Bruno Grisci[2], Rodrigo Ligabue-Braun[1], and Marcio Dorn[2*]
[1] *Institute of Biosciences, Federal University of Rio Grande do Sul*
[2] *Institute of Informatics, Federal University of Rio Grande do Sul*
[*] mdorn@inf.ufrgs.br

Proteins are the most abundant biological macromolecules in nature, they occur in great variety. Thousands of different proteins can be found in a single cell, working as mediators of each one of the processes taking place in the cells. The protein's structure can be defined as a conceptual hierarchy: the primary structure describes the linear sequence of amino acids residues while the secondary structure refers to spatial arrangements that follow some regular patterns. The tertiary structure refers to the 3D folding of a protein and represents the functional state of the molecule.

Single nucleotide polymorphism (SNP) occurs very commonly throughout the genome, these changes can result in no change in amino acid sequence, called synonymous SNP (sSNP), or lead to amino acid changes, called non-synonymous SNP (nsSNP) [RBS02]. SNPs are significant research objects once they are directly associated with pathologies and are used as genetic markers.

Predicting the effects of SNPs is a challenge and still an open research field, for which there are two main approaches: experimentally, which is expensive and susceptible to errors due to the wrong manipulation, and computationally, which is a lot faster and cheaper. The available computational approaches differ on which parameters they choose to increase the accuracy, but only a few of them use the secondary structure as a parameter. The 3D structure of proteins follows a hierarchy; thus a change of amino acid residue in the peptide sequence could cause the rupture of secondary structures, hindering protein's folding, resulting in functional changes.

The aim of this study is to develop a new tool (namely, ROMA), using non-usual approaches, to predict the effects of SNPs using information about the secondary structure as parameters. This parameter will be drawn up from local conformational preference retrieved from a database when there is a replacement of a single amino acid residue in the amino acid sequence.

A search for proteins that presents some functional change due to a substitution of a single amino acid residue was executed through databases that contain information about SNPs, then the secondary structure of this residue and its left and right neighbors were defined. We developed a method based on the analysis of dihedral angles, phi, and psi, to count the occurrence of this angles combination occurring on RCSB PDB [BWF+06], for each amino acid replacement and its neighbors.

The distances between the distributions were calculated using the *Earth Movers Distance* (EMD) algorithm so that we can rank the bests candidates. To validate our findings, *Molecular Dynamics simulations* were performed using the GROMACS *package*. The validation was done by crossing the results from EMD analysis and GROMACS simulations, where the best candidates are expected to behave the same structural proprieties as the non-mutated protein over the simulation. So far the obtained results are not conclusive; however, they're promising, but due to the time needed to perform the molecular dynamics we are still running more test cases.

References

[BWF+06] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank, 1999–. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, pages 675–684. Springer, 2006.

[RBS02] Vasily Ramensky, Peer Bork, and Shamil Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic acids research*, 30(17):3894–3900, 2002.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

A novel method to identify viral microRNAs combining local features and comparative genomics

Kevin Lamkiewicz[1,*] and Manja Marz[1,2]
[1] *RNA Bioinformatics and High Throughput Analysis, Friedrich Schiller University Jena, Germany*
[2] *European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany*
[*]kevin.lamkiewicz@uni-jena.de

In the last years, several microRNAs (miRNAs) encoded by viruses were identified [PZG+04]. Hitherto, there are 500 viral miRNAs distributed among 29 virus species submitted to the *mirbase.org* database [KGJ14]. Tools for miRNA prediction can be categorized into two groups: homology-based approaches and machine learning approaches. Both strategies have been applied successfully in eukaryotes. However, none of the approaches is completely applicable for viruses as requirements are usually not met. Homology-based methods require knowledge of phylogenetic close species, whereas machine learning methods require a sufficient training set.

We propose the tool `ViMiFi` (**vi**ral **mi**RNA **fi**nder) that is able to predict precursor miRNAs in both single sequences and multiple sequence alignments. Several established models for eukaryotic miRNA prediction based on support vector machines were modified and used to train a Random Forest Classifier (RFC). `ViMiFi` is able to either scan a single sequence for potential precursor miRNAs or to analyze regions of a multiple sequence alignment that are conserved based on their secondary structure. The machine learning based classifier is strongly inspired by `TripletSVM` [XLH+05] that introduces triplet-features. These triplets model local sequence and structural features of the sequence. We combined these triplet features with other common features, such as thermodynamic stability, secondary structure and base-pairing features.

Our RFC achieves a high sensitivity and specificity with three different negative training sets. With our method, we detected the majority of already known miRNAs and identified novel miRNA candidates throughout several virus families. Furthermore, we compared the training results of our classifier with `MiRenSVM` [DZG10] and `TripletSVM`. Our analysis shows that `ViMiFi` achieves a higher sensitivity and specificity for viral data.

Emerging data indicate viral miRNAs to either regulate viral replication or suppress counter mechanisms in infected cells. Therefore, a comprehensive overview of miRNAs in RNA viruses might lead to a general approach to disturb viral replication and facilitate the development of new medical therapies.

References

[DZG10] Jiandong Ding, Shuigeng Zhou, and Jihong Guan. MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. *BMC bioinformatics*, 11 Suppl 11:S11, December 2010.

[KGJ14] Ana Kozomara and Sam Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 42(Database issue):D68–D73, 2014.

[PZG+04] Sébastien Pfeffer, Mihaela Zavolan, Friedrich A Grässer, Minchen Chien, James J Russo, Jingyue Ju, Bino John, Anton J Enright, Debora Marks, Chris Sander, and Thomas Tuschl. Identification of virus-encoded microRNAs. *Science (New York, N.Y.)*, 304:734–736, April 2004.

[XLH+05] Chenghai Xue, Fei Li, Tao He, Guo-Ping Liu, Yanda Li, and Xuegong Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinf*, 6:310, December 2005.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Rapid Estimation of Evolutionary Distances between Bacterial Genomes

Fabian Klötzl and Bernhard Haubold
Max Planck Institute for Evolutionary Biology, Plön
kloetzl@evolbio.mpg.de

The estimation of phylogenies from distance matrices is dominated by computations that are quadratic in the number of taxa analysed: Traditionally, all pairs of sequences need to be aligned, and the distance computed for each aligned pair. A while ago, we devised a method for merging these two steps through the computation of anchor distances [HKP15]. Our implementation ANDI returns fast and accurate distances between large samples of whole bacterial genomes. Still, the runtime of ANDI is dominated by the pairwise search for exact matches that underlies the construction of anchor distances. To improve on this, we are currently working on a method where the input sequences are not compared to each other, but stacked onto a single reference sequence. The resulting stacks of homologous regions correspond to multiple local alignments, from which the distances can be computed as they would be from a traditional multiple sequence alignment. Due to the linear number of matches involved, this stack-based method is much faster than the traditional computation of anchor distances. We demonstrate the performance of this new approach by applying it to all *Escherichia coli* genomes from the ENSEMBL database.

References

[HKP15] Bernhard Haubold, Fabian Klötzl, and Peter Pfaffelhuber. andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31(8):1169, 2015.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Understanding global patterns of structural variation in Arabidopsis thaliana - Future of the 1001 Genome project

Felix Bemm, The 1001 Genomes Consortium
Max Planck Institute for Developmental Biology, Tuebingen
felix.bemm@tuebingen.mpg.de

The recently released map of polymorphism of 1,135 re-sequenced *A. thaliana* natural inbred lines provides an invaluable resource to understand global genetic patterns in a large collection of wild individuals that are products of natural selection under diverse ecological conditions.

Our work clarified prior hypotheses such as the strong impact of the last ice age on population structure and in addition revealed that modern *A. thaliana* is a mixture of stationary relicts and fast expanding survivors from different glacial refugia. In the future the resource will enable researchers to decipher more accurately how genetic variation translates into phenotypic variation.

However, connecting hits from genome-wide association studies (GWAS) to causal sequences is still challenging since previous analysis approaches mostly rely on mapping to a single reference, which insufficiently captures structural variations (SVs) or the presence of sequences not found in the reference. Discovery and genotyping of such sequences remains computationally difficult with short read data, since they are often in repetitive regions and because they can the changes can be complex.

The next objective for the 1001 Genomes project is to discover and genotype major classes of SVs in the global set of *A. thaliana* accessions. By reanalyzing the 1001 Genomes Project whole-genome sequencing (WGS) data together with long-read DNA technologies, we will survey SV mutation hotspots throughout the worldwide population and target so far undescribed patterns and classes of SV complexity.

Here we present strategies, pilot studies and first results of our upcoming work that will provide structural variations and polymorphisms as an integrated resource alongside with detailed information about epigenomes as well as molecular and nonmolecular phenotypes to understand how traits are connected to genomes and epigenomes.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

MetaMeta: Integrating metagenome analysis tools to improve taxonomic profiling

Vitor C. Piro and Marcel Matschkowski and Bernhard Y. Renard
Robert Koch Institute - Research Group Bioinformatics (MF1), Nordufer 20, 13353, Berlin, Germany
PiroV@rki.de, MatschkowskiM@rki.de, RenardB@rki.de

Many metagenome analysis tools are presently available to classify sequences and profile environmental samples. In particular, taxonomic profiling and binning methods are commonly used for such tasks. Tools available among these two categories make use of several techniques, e.g. read mapping, k-mer alignment, and composition analysis. Variations on the construction of the corresponding reference sequence databases are also common. In addition, different tools provide good results in different datasets and configurations. All this variation creates a complicated scenario to researchers to decide which methods to use. Installation, configuration and execution can also be difficult especially when dealing with multiple datasets and tools.

We propose MetaMeta: a pipeline to execute and integrate results from metagenome analysis tools. MetaMeta provides an easy workflow to run multiple tools with multiple samples, producing a single enhanced output profile for each sample. MetaMeta includes a database generation, pre-processing, execution, and integration steps, allowing easy execution and parallelization. The integration relies on the co-occurrence of organisms from different methods as the main feature to improve community profiling while accounting for differences in their databases.

In a controlled case with simulated and real data we show that the integrated profiles of MetaMeta overcome the best single profile. Using the same input data, it provides more sensitive and reliable results with the presence of each organism being supported by several methods. MetaMeta uses Snakemake and has six pre-configured tools, all available at BioConda channel for easy installation (conda install -c bioconda metameta). The MetaMeta pipeline is open-source and can be downloaded at: https://gitlab.com/rki_bioinformatics

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Tools for the Management of Illumina Flowcells and Demultiplexing of Raw Base Calls

Manuel Holtgrewe[1,2] and Dieter Beule[1,3]
[1]*Berlin Institute of Health,* [2]*Charité Universitätsmedizin Berlin,* [3]*Max Delbrück Center for Molecular Medicine*
manuel.holtgrewe@bihealth.de

While mundane work, the metadata management of flowcells for Illumina sequencers as well as the demultiplexing of raw base call files is still fundamental to all subsequent analyses. A number of facilities are employing a laboratory information system (LIMS) for the management of their samples and flow cells. However, because of the high cost of establishing commercial or developing custom solutions, many labs still stick to a "Spreadsheets on network shares" solution. While the development of a LIMS system is out scope for the presented work, we address the two issues of (1) creating sample sheets for the Illumina demultiplexing software and (2) performing the demultiplexing and subsequent data QC in a reproducible and automated step.

First, flowcelltool is a web app developed with Python and Django that allows for the easy management of flow cells and libraries. The software also manages a library of adapter sequences, such that these can be conveniently selected by their names instead of entering their names. The software is easy to use, allows the easy import of sample sheet information from Spreadsheets by copy and pase, provides basic search functionality and also deals with automatically reverse-complementing adapters when necessary for dual indexing setups. Finally, the software can generate sample sheets for direct use with the Illumina bcl2fastq (v1 or v2) software or a YAML-based version for cubi_demux.

Second, cubi_demux is a command line tool based on Snakemake [KR12]. Given a folder with the sequencing run output and a sample sheet (as generated by Flowcelltool), cubi_demux runs demultiplexing and subsequent data QC automatically. The correct version of bcl2fastq is called transparently, depending on the RTA version used for sequencing, and the resulting data is subjected to FastQC [And10], a screening for containing model organism sequence. Any read not mapping to a model organism reference is then subjected to a metagenomics screening using Kraken [WS14].

The software is available under permissive open source licenses at https://github.com/bihealth/flowcelltool and https://github.com/bihealth/cubi_demux. We welcome contributions and discussions with users.

References

[And10] S. Andrews. FASTQC. A quality control tool for high throughput sequence data, 2010.

[KR12] J. Köster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, Oct 2012.

[WS14] D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, 15(3):R46, Mar 2014.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Onctopus: Subclonal Reconstruction of Cancer Samples based on SNVs and CNAs

Linda K. Sundermann[1,2], Amit G. Deshwar[3,4], Jeff Wintersinger[5], Daniel Doerr[2], Quaid Morris[3,6], and Gunnar Rätsch[7,8]
[1] *International Research Training Group GRK 1906/1 "Computational Methods for the Analysis of the Diversity and Dynamics of Genomes"*, [2] *Genome Informatics, Faculty of Technology, and Institute for Bioinformatics, Center for Biotechnology, Bielefeld University, Germany,* [3] *Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Canada,* [4] *Deep Genomics Inc., Canada,* [5] *Department of Computer Science, University of Toronto, Canada,* [6] *The Donnelly Center for Cellular and Biomolecular Research, University of Toronto, Canada,* [7] *Biomedical Informatics, Department of Computer Science, ETH Zürich, Switzerland,* [8] *Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, USA*
lsunderm@cebitec.uni-bielefeld.de, amit.deshwar@utoronto.ca, jeff@wintersinger.org, ddoerr@cebitec.uni-bielefeld.de, quaid.morris@utoronto.ca, gunnar.ratsch@ratschlab.org

Cancer samples are often genetically heterogeneous, harboring subclonal populations (subpopulations) with different mutations such as copy number aberrations (CNAs) or single nucleotide variants (SNVs). Information about such mutations in the subpopulations can help to identify driver mutations or to choose targeted therapies. Sequencing of bulk tumor samples is current standard practice because single-cell assays are not yet well established due to high cost and limited resolution.

Recently, several methods that attempt to infer the genotype of subpopulations using CNAs, SNVs, or both [DVY+15, JQMZ16] have been published. Our new approach *Onctopus* also utilizes CNAs and SNVs and models them jointly to reconstruct the subclonal composition of a bulk tumor sample.

Given haploid average copy numbers of segments affected by CNAs and variant counts of SNVs, *Onctopus* assigns a frequency, CNAs and SNVs to N subclonal lineages. Each of these lineages is defined through the CNAs and SNVs that arose in this lineage.

We build a joint likelihood model and model the tumor as consisting of a mixture of lineages. We choose subclonal lineages to avoid ambiguous solutions that can occur when copy numbers are determined for subpopulations. We developed a linear relaxation of our model as a mixed integer linear program that can be solved with state-of-the-art solvers.

References

[DVY+15] Amit G. Deshwar, Shankar Vembu, Christina K. Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015.

[JQMZ16] Yuchao Jiang, Yu Qiu, Andy J. Minn, and Nancy R. Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37):E5528–E5537, 2016.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Genome-wide Analysis of Small RNA-controlled Gene networks in Maize Leaf Development

Xiaoli Ma, Marie Javelle, Steffen Knauer, Patrick Schnable, Jianming Yu, Gary Muehlbauer, Mike Scanlon and Marja C. P. Timmermans
University of Tbingen
xiaoli.ma@uni-tuebingen.de

In plants, stem cell niches serve as a stable source of cells for postembryonic growth and development. The shoot apical meristem (SAM) gives rise to all aerial organs of a plant, and its activity throughout the plants lifetime therefore has to be tightly controlled in a spatiotemporal manner. To gain insight into gene regulatory networks behind stem cell maintenance and organogenesis, we generated a high-resolution gene expression atlas of 12 distinct domains within the vegetative maize shoot apex using laser microdissection and RNA deep sequencing. We also generated small RNA sequencing data that informs on the role of miRNAs in the maize shoot apex. Together these data reveal a subfunctionalization of miRNA family members across the SAM subdomains, and the regulation of miRNA accumulation in the stem cell containing SAM tip. In addition, miRNA degredome sequencing data were produced, combined with information from the SAM atlas, we predicts the presence of mechanisms that further fine-tune the accumulation and activity of select small RNAs to regulate key meristem genes.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

DACCOR - Detection, charACterization, and reCOnstruction of Repetitive regions in bacterial genomes

Alexander Seitz, Friederike Hanssen, and Kay Nieselt
Center for Bioinformatics (ZBIT), Integrative transcriptomics, Eberhard-Karls-Universität Tübingen
alexander.seitz@uni-tuebingen.de

The reconstruction of genomes using mapping based approaches with short reads experiences difficulties when resolving repetitive regions. Reads that cannot be placed at a unique position are assigned low mapping qualities [LRD08]. However, reads that stem from closely related species are also assigned low quality scores, which is why in the field of ancient DNA (aDNA), where we always deal with a metagenomic sample, these reads are often filtered out [B+16]. If not filtered out, these low mapping qualities lead to low genotyping qualities [M+10] and thus often unresolved bases [SXZ08]. A typical approach to address repetitiveness is to use reference genomes for mapping, in which known repetitive regions are masked [FHH10]. However, for many references such masked genomes are not available or are based on repetitive regions of other genomes [TGC09]. Here, we present a combined approach, which first identifies repeats *de novo* in a given reference genome. These regions can then be used to reconstruct them separately using short read sequencing data. Afterwards, the assembled repetitive sequences can be inserted into the reconstructed genome. We present the program `DACCOR`, which performs these steps automatically. Our results for several bacterial genomes show an increased base pair resolution of the repetitive regions, in comparison to standard mapping based assembly approaches.

References

[B+16] Kirsten I. Bos et al. Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *eLife*, 5:1–11, 2016.

[FHH10] Martin C Frith, Michiaki Hamada, and Paul Horton. Parameters for accurate genome alignment. *BMC bioinformatics*, 11(1):80, 2010.

[LRD08] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, 2008.

[M+10] Aaron McKenna et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[SXZ08] Andrew D Smith, Zhenyu Xuan, and Michael Q Zhang. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, 9(1):128, 2008.

[TGC09] Maja Tarailo-Graovac and Nansheng Chen. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, pages 1–14, 2009.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

German Network for Bioinformatics Infrastructure

F. Sprengel, T. Dammann-Kalinowski, D. Jording, I. Maus, A. Tauch, D. Wibberg, A. Pühler
de.NBI Administation Office, c/o Center for Biotechnology, Bielefeld University
contact@denbi.de

In recent years, the modern life sciences research underwent a rapid development that was driven mainly by the technical improvements in analytical areas in terms of miniaturization, parallelization and high through-put of biological samples and thus the generation of huge amounts of experimental data. Prominent examples of this ongoing development are the omics techniques featuring the analysis of the various levels of information storage and processes in living cells, and the numerous new imaging techniques providing insights into biological systems to a hitherto unprecedented depth. The ever growing application of these novel techniques and the exploitation of the resulting data have revolutionized many fields of science and are furthermore opening new areas of basic and applied research with considerable opportunities for life sciences. The bottleneck that prevents realization of the full potential of the different omics technologies is not the data generation itself, but the subsequent data analysis.

The German Network for Bioinformatics Infrastructure (de.NBI) takes care of this challenge in many areas of life sciences with its mission to provide, expand and improve a repertoire of specialized bioinformatics tools, appropriate computing and storage capacities and high-quality data resources. These efforts are supplemented by a training program providing courses on the supplied tools.

de.NBI is a distributed bioinformatics infrastructure which started in March 2015 as an academic funding initiatve of the German Ministry of Research and Education (BMBF). The consortium currently consists of 39 project partners organised in eight service centers and one central administration and coordination unit. The service centers offer a variety of training courses and bioinformatics services, online databases, software libraries, and tools as webservices and/or for download. Furthermore, consulting on individual issues is available. Services are aimed at application users in life sciences as well as bioinformaticians and developers. The de.NBI services will be unified with regard to standards, interchangeability and reproducibility. The network has recently been supplemented with a federated cloud at five locations. This hardware is enabling big data exploitation in all areas of life sciences.

Libraries for Variants and Phenotypes in Clinical Applications

Manuel Holtgrewe[1,2], Max Schubach[1], Sebastian Köhler[2], Dieter Beule[1,3], and Peter N. Robinson4
[1]*Berlin Institute of Health,* [2]*Charité Universitätsmedizin Berlin,* [3]*Max Delbrück Center for Molecular Medicine,* [4]*The Jackson Laboratory*
manuel.holtgrewe@bihealth.de

While high-throughput sequencing for genetic screening has become commonplace in clinical applications, the interpretation of variants from such sequencing still poses a great challenge. In particular, as efficient computational methods for the calling of germline variants have been established, the amount of data in clinical contexts is expected to rise. Already today, the interpretation of variants by physicians and counselors poses the bottleneck for wide-spread adoptions of clinical or whole exome sequencing.

To facilitate the development of future tools, we present our library Ontolib and updates to our library (and accompanying application) Jannovar. This software has been used in previous work such as the Phenomizer [**?**] and Phenix [Phe] and was recently also included in the Exomiser [**?**]. Now it is also readily available for other authors to use in the form of libraries.

Ontolib is a Java library for representing biological ontologies for phenotypical similarity search as well as important algorithms for working with ontologies such as the human phenotype ontology (HPO), gene ontology (GO), and mammalian phenotype ontology (MPO).

Jannovar [**?**] is a software package consisting of a Java library and program for the annotation of variants with predicted molecular impact. The updates include the full support of VCF and the annotation with frequency information and scores from the most important data bases, including Clinvar, dbSNP, ExAC, gnomAD, dbNSFP etc. Further, Jannovar contains code for parsing HGVS variant descriptions and representing them as Java objects as well as annotation variant files with compatible mode of inheritance.

All software is available under permissive open source licenses at https://github.com/charite/jannovar and https://github.com/phenomics/ontolib and we welcome contributions by and discussions with users.

References

[JWB$^+$14] M. Jäger, K. Wang, S. Bauer, D. Smedley, P. Krawitz, and P. N. Robinson. Jannovar: a java library for exome annotation. *Hum. Mutat.*, 35(5):548–555, May 2014.

[KSK$^+$09] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dolken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.*, 85(4):457–464, Oct 2009.

[SJJ$^+$15] D. Smedley, J. O. Jacobsen, M. Jäger, S. Köhler, M. Holtgrewe, M. Schubach, E. Siragusa, T. Zemojtel, O. J. Buske, N. L. Washington, W. P. Bone, M. A. Haendel, and P. N. Robinson. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*, 10(12):2004–2015, Dec 2015.

[ZKM$^+$14] T. Zemojtel, S. Köhler, L. Mackenroth, M. Jäger, J. Hecht, P. Krawitz, L. Graul-Neumann, S. Doelken, N. Ehmke, M. Spielmann, N. C. Oien, M. R. Schweiger, U. Krüger, G. Frommer, B. Fischer, U. Kornak, R. Flottmann, A. Ardeshirdavani, Y. Moreau, S. E. Lewis, M. Haendel, D. Smedley, D. Horn, S. Mundlos, and P. N. Robinson. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med*, 6(252):252ra123, Sep 2014.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Reliability of data detection and processing approaches in chromatography

David Thomas Marehn[1,2] and Heike Pospisil[1]
[1] *UAS Wildau, High Performance Computing in Life Sciences*
[2] *Università degli Studi di Roma "Tor Vergata"*
pospisil@th-wildau.de

Nowadays pharmaceutical analysis and industry couldn't be imagined without using chromatographic methods like the High Performance Liquid Chromatography (HPLC) and Gas Chromatography (GC). Therefore, the field of chromatography is already firmly anchored in the three important, regional pharmacopeias: EP, USP and JP. These pharmacopeias list the types, calculable data and suitable parameters of chromatographic methods. On closer examination of the chapter dealing with the parameters the detector module isn't as well defined as expected because it seems that two configurable parameters are missing. Several present investigations have shown that the sampling rate and the filtering of the acquired data influence the following data processing.

A still unsolved problem is the definition of optimal settings for the chromatographic engine. These setting parameters (the detector data rate and the filtering) depend on the currently used detectors and manufacturer specific chromatographic data system (CDS) software packages and are not known per se.

A newly developed software allows the acquisition of data from the HPLC detector parallel to the controlling CDS. Two HPLC systems and two CDS using a well defined sample standard have been tested. For the data processing a modification of an existing integration algorithm has been used. It includes peak detection and integration without any unwanted smoothing of the data. A statistical analysis of the evaluated chromatograms shows that there are no significant deviations between the acquired data of the CDS compared to the own written data collector. Based on the fact that the data collector does not modify the incoming data this also applies to the behavior of the two tested CDS.

The gained experience by the experiments will help to develop algorithms in order to find the optimal sample rate and filtering settings for the best data processing. Further experiments will be done using an external A/D converter due to the available fine adjustment of the data acquisition comparing to the internal A/D converter of the detector module.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Quality Control of Affymetrix GeneChips - Automatic Detection of Spatial Patterns in Pseudo Images by AffyQCImage

Yang Xiang[1], Mila Vukmirovic[2,1], and Florian Martin[1]
1. Philip Morris International Research and Development, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchatel, Switzerland (part of Philip Morris International group of companies)
2. Section de mathematiques, EPFL FSB SMA, Station 8 - Batiment MA, CH-1015 Lausanne, Switzerland
Yang.Xiang@pmi.com, Florian.Martin@pmi.com

Affymetrix GeneChips are widely used for gene expression profiling. Data quality control (QC) is crucial to ensure robust and reliable biological conclusions. In the R statistical environment, several packages support the automated QC of Affymetrix GeneChips using metrics such as the Normalized Unscaled Standard Error (NUSE) and the Relative Log Expression (RLE) [WM05, BCB+05, KGH08]. Pseudo images are also widely used to detect spatial artifacts on individual GeneChips. However, no automated procedure for the evaluation of pseudo images is currently available, which limits the objectivity and reproducibility of this QC metric. Here, we present AffyQCImage, an R package for the automated detection of spatial patterns in pseudo images that can complement the other available QC metrics for Affymetrix GeneChips. Our algorithm automatically denoises the data using both Markov random fields and generalized additive models. Subsequently, for each GeneChip, the denoised data are one-dimensionally clustered into background and outlier spots. Finally, based on the outlier-background distance and outlier area, GeneChips of poor quality are identified. We tested this algorithm with manually derived training/testing data from 6517 GeneChips using 5-fold cross validation with 20 repeats. The mean accuracy and mean Matthews correlation coefficient (MCC) were 0.90 and 0.67 for human (HG-U133_Plus_2), 0.94 and 0.77 for mouse (Mouse430_2), and 0.98 and 0.80 for rat (Rat230_2) GeneChips. This robust performance suggests that the evaluation of pseudo images using the AffyQCImage package can be utilized in automated QC pipelines for Affymetrix GeneChip data.

References

[BCB+05] B Bolstad, François Collin, Julia Brettschneider, K Simpson, L Cope, R Irizarry, and Terence P Speed. Quality assessment of Affymetrix GeneChip data. *Bioinformatics and computational biology solutions using R and bioconductor*, pages 33–47, 2005.

[KGH08] Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayQualityMetricsa bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416, 2008.

[WM05] Claire L Wilson and Crispin J Miller. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics*, 21(18):3683–3685, 2005.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Transcriptome Analysis in Hybrid Plants using Platinum Genomes

Maximilian Collenberg, Felix Bemm & Detlef Weigel
Max Planck Institute for Developmental Biology, Tübingen
max.collenberg@tuebingen.mpg.de

Heterosis refers to the deviation of the F1 progeny from the phenotypic mean of the parental plants, especially in cases where both parents are inbred and homozygous throughout the genome. When comparing certain traits of a F1 hybrid to the corresponding parental plants one can compare the performance of the hybrid to (1) the midparent value (MPV), which refers to the mean of both parental plants for the given trait. Furthermore the F1 hybrid can be compared to (2) the best-parent value (BPV), referring to the value of the superior parental plant. Mainly three genetic models, explaining heterosis have been proposed. However, there is no consensus about the diversity of molecular principles of heterosis. Recent studies suggested structural genome variation, such as copy-number variation (CNV) as well as presence absence variation (PAV), among maize inbred lines, underlying complementary contributions of genes from both parents as an important factor (Springer et al. 2009). Gene expression with regard to heterosis has been analyzed in several plants including*A. thaliana* (Alonso-Peral et al. 2017). However, the majority of these studies relied either on microarray data, which are limited to transcripts present on the chip, or on RNA-seq data that have been processed using a single reference genome, thus not accounting for structural variation among different genotypes (accessions). We now have access to high quality full length genome sequences of different *A. thaliana* accessions. Two of these accessions have been used in a reciprocal crossing. The root and shoot transcriptomes of F1 hybrids as well as from the corresponding parental plants have been short read sequenced. RNA-seq reads have been processed using either one parental genome as a reference or a trans-reference genome, containing full length genome sequences of both parental plants and accounting for orthologous assignment. DESeq2 has been used to detect transcripts with non parental expression patterns (below or above MPV) in the F1 hybrid plants. When using a single reference genome we found various transposable elements among the differentially expressed genes ($p < 0.01$). However, whole genome alignment data among parental plants indicate that many of these are due to a reference bias. Here we present a way of how to process a double reference genome in order to improve the analysis of hybrid transcriptomes regarding heterosis.

References

Alonso-Peral, Maria M et al. (2017). “Patterns of gene expression in developing embryos of Arabidopsis hybrids”. en. In: *Plant J.* 89.5, pp. 927–939.

Springer, Nathan M et al. (2009). “Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content”. en. In: *PLoS Genet.* 5.11, e1000734.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

DYNAMITE - A discriminatory method for the identification of key transcriptional regulators using epigenetic data

Florian Schmidt [1,2,3], Tzung-Chien Hsieh [4], Nina Gasparoni [5], Gilles Gasparoni [5], Julia K. Polansky [6], Oliver Gorka [7], Jürgen Ruland [7], Karl Nordström [5], Anupam Sinha [8], Wei Chen [9], Alf Hamann [6], Philip Rosenstiel [8], Jörn Walter [5], and Marcel H. Schulz [1,2,*]

[1]Cluster of Excellence for Multimodal Computing and Interaction, Saarland Informatics Campus, Saarbrücken, Germany. [2]Computational Biology & Applied Algorithmics, Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany. [3]Graduate School for Computer Science, Saarland Informatics Campus, Saarbrücken, Germany. [4]Charité - Universitätsmedizin Berlin, Berlin, Germany. [5]Department of Genetics, Saarland University, Saarbrücken, Germany. [6]Experimental Rheumatology, German Rheumatism Research Centre, Berlin, Germany. [7]Institute for Clinical Chemistry and Pathobiochemistry, Klinikum rechts der Isar, Technical University Munich, Germany. [8]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. [9]Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany.
fschmidt@mmci.uni-saarland.de

Deciphering the regulatory mechanism that control the establishment and maintenance of cellular programs is an essential task in computational biology. Transcription Factors (TFs) are key players in these mechanisms. An established approach to understand how TFs regulate changes in gene expression between tissues is through TF-ChIP-seq experiments. While accurate, these experiments are laborious, time-consuming, and need a priori knowledge of relevant TFs. Alternative approaches, which combine solely sequence-based TF-binding predictions with TF gene expression measurements, are less accurate in describing both gene expression differences and TF-target relationships.

Recently, a number of genome-wide open-chromatin assays, e.g. DNaseI-seq, have been utilized to measure chromatin accessibility in a sample of interest. These measurements can be combined with computational TF-binding annotations [S$^+$17]. Here, we present *DYNAMITE*, a two-step machine learning approach using only open-chromatin data as input to identify TFs that might be key regulators of gene expression differences between tissues. First, a new statistical approach is developed for the computation of differential TF-binding scores for each gene and TF, using binding predictions computed in open-chromatin regions, incorporating open-chromatin replicate information, and weighted binding in far-away enhancer regions. Second, an interpretable logistic regression classifier is used to prioritize TFs that are best suited to discriminate up and down regulated genes. *DYNAMITE* outperforms purely sequence-based approaches, and performs comparable to methods based on TF-ChIP-seq data [C$^+$12].

As part of DEEP and IHEC, we have applied our model to identify TFs that are key regulators of human CD4+ T cell differentiation from naive (TN) to effector memory T cells (TEMs). *DYNAMITE* highlighted several TFs as regulators of these transitions, including FOXP1, which was suggested to discriminate TNs from TEMs. This prediction was validated experimentally: T cell-specific depletion of FOXP1 in knock-out mice indeed resulted in loss of the naive T-cell phenotype [D$^+$16].

Overall, we suggest an accurate and flexible method to identify key regulatory TFs of gene expression between tissues. As only open-chromatin and gene expression data are required, comparatively low experimental costs allow our method to be applied to many cellular systems. The complete workflow is available in a user-friendly GUI and in bash scripts (*https://github.com/SchulzLab/TEPIC*).

References

[C$^+$12] C. Cheng et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, 22(9):1658–1667, Sep 2012.

[D$^+$16] P. Durek et al. Epigenomic Profiling of Human CD4(+) T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. *Immunity*, 45(5):1148–1161, Nov 2016.

[S$^+$17] F. Schmidt et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.*, 45(1):54–66, Jan 2017.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

How to analyze raw sequencing data from microbiota? Using ASaiM!

Berenice Batut[1], Clemence Defois[2], Kevin Gravouil[2], Jean-Francois Brugere[2], Eric Peyretaillade[2], Pierre Peyret[2]
[1] *University of Freiburg, Germany*
[2] *MEDIS/LMGE/LIMOS, Universit Clermont Auvergne, France*
berenice.batut@gmail.com

The study of microbiota and microbial communities has been facilitated by the evolution of sequencing techniques and the development of metagenomics and metatranscriptomics. These techniques are giving insight into phylogenetic properties and metabolic components of microbial communities. However, meta'omic data exploitation is not trivial: large amount of data, high variability, incompleteness of reference databases, difficulty to find, configure, use and combine the dedicated bioinformatics tools, etc. Hence, to extract useful information, a sequenced microbiota sample has to be processed by sophisticated workflows with numerous successive bioinformatics steps. Besides, bioinformatics tools are often manually executed and/or patched together with custom scripts. These practices raise doubts about a science gold standard: reproducibility. Alternative approaches to improve accessibility, modularity and reproducibility can be found in Open-Source workflow systems such as Galaxy. Galaxy is a lightweight environment providing a web-based, intuitive and accessible user interface to command-line tools, while automatically managing computation and transparently managing data provenance and workflow scheduling [68]. In this context, we developed ASaiM (Auvergne Sequence analysis of intestinal Microbiota), an Open-Source opinionated Galaxy-based framework.

ASaiM provides an expertly selected collection of tools to exploit and visualize taxonomic and functional information from raw amplicon, metagenomic or metatranscriptomic sequences. To help the analyses, several (customizable) workflows are included. The main workflow has been tested on two mock metagenomic datasets with controlled communities. More accurate and precise taxonomic analyses and more informative metabolic description have been obtained compared to EBI metagenomics' pipeline on the same datasets.

The available workflows are supported by tutorials and Galaxy interactive tours to guide the users through the analyses. Furthermore, an effort on documentation of ASaiM, its tools and workflows has been made (http://asaim.readthedocs.io/).

Based on the Galaxy framework, ASaiM offers sophisticated analyses to scientists without command-line knowledge, while emphasizing reproducibility, customization and effortless scale up to larger infrastructures. ASaiM is implemented as Galaxy Docker flavour and can be easily extended with additional tools or workflows. ASaiM provides then a powerful framework to easily and quickly exploit microbiota data in a reproducible and transparent environment.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Exploded Views for Protein Structures

Mirjam Figaschewski, Julian Heinrich
Applied Bioinformatics Group, University of Tuebingen
figasch@informatik.uni-tuebingen.de

Exploded views are a common technique in engineering to visually convey the composition of complex objects. We developed a fully automatic approach for the generation of exploded views for protein complexes as a plugin for PyMOL [DeL02].

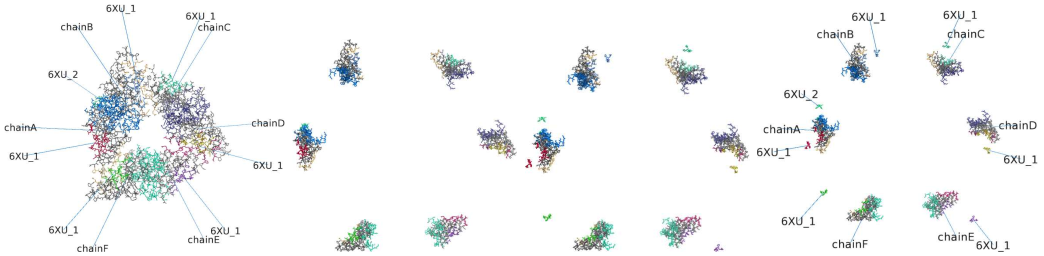
In this illustration technique the protein is segmented into chains and ligands and these segments are displaced hierarchically to expose occluded details, e.g., ligands that are buried in the structure. The steps of an explosion are shown exemplarily in Figure 1. Rather than a static exploded view, a movie of the explosion is generated here.

An explosion direction has to be chosen, so that the segments do not overlap in the resulting view. The implemented explosion direction of a segment is a vector defined by the *Center Of Mass* (*COM*) of the whole structure and the *COM* of the segment. All segments explode from the same center with the same distance to their original position.

To ensure that the segments of an exploded view can be identified easily by the viewer, the segments are labelled. A *radial circular* layout has been chosen. The line, connecting a label with its related segment, is equivalent to the explosion vector of the segments.

Additionally, the contact sites between chains and the binding sites can be highlighted by different coloring: *contact-based* (Figure 1) or *chain-based*.

The explosion movie is interactive, easy to use and assists in an intuitive interpretation of the process. By using the explosion movie, the participants of an empirical study identified chains, ligands and contacts between them about three times faster than without it.



(a) Original complex with labels. (b) Chains exploded. (c) Ligands exploded. (d) Final exploded view.

Figure 1: Explosion of a protein, colored contact-based: every contact area between chains or between chains and ligands has a unique color.

References

[DeL02] W DeLano. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter On Protein Crystallography*, 700, 2002.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Real time pathogen identification from metagenomic Illumina datasets

Simon H. Tausch, Jakob Schulze, Andreas Andrusch, Tobias P. Loka, Jeanette Klenner, Piotr W. Dabrowski, Bernhard Y. Renard, Andreas Nitsche
Robert Koch Institute, Berlin, Germany
tauschs@rki.de

In the past years, Next Generation Sequencing has been utilized in time critical applications such as pathogen diagnostics with promising results. Yet, long turnaround times had to be accepted to generate sufficient data, as the analysis was performed sequentially after the sequencing was finished. Finally, the interpretation of results can be hindered by various types of contaminations, clinically irrelevant sequences, and the sheer amount and complexity of the data. We designed and implemented a real-time diagnostics pipeline which allows the detection of pathogens from clinical samples up to five days before the sequencing procedure is even finished. To achieve this, we adapted the core algorithm of HiLive, a real-time read mapper, while enhancing its accuracy for our use case. Furthermore, common contaminations, low-entropy areas, and sequences of widespread, non-pathogenic organisms are automatically marked beforehand using NGS datasets from healthy humans as a baseline. The results are visualized in an interactive taxonomic tree, providing the user with several measures regarding the relevance of each identified potential pathogen. We applied the pipeline on a human plasma sample spiked with Vaccinia virus, Yellow fever virus, Mumps virus, Rift Valley fever virus, Adenovirus and Mammalian orthoreovirus, which was then sequenced on an Illumina HiSeq. All spiked agents could already be detected after only 12% of the complete sequencing procedure. While we also found a large number of other sequences, these are correctly marked as clinically irrelevant in the resulting visualization, allowing the user to obtain the correct assessment of the situation at first glance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Implementing a multilayer framework for pathway data integration, analysis and visualization

Zaynab Hammoud and Frank Kramer
Universitätsmedizin Göttingen, Institut fr Medizinische Statistik
zaynab.hammoud@med.uni-goettingen.de

Personalized medicine, i.e. a medicine focused on the individual and proactive in nature, promises an improved health care by customizing the treatment according to patient needs [B+14]. The methods to analyse data, model knowledge and store interpretable results vary widely. A common approach is to use networks for modelling and organizing this information. Network theory has been used for many years in the modelling and analysis of complex systems, as epidemiology, biology and biomedicine [k+14]. As the data evolves and becomes more heterogeneous and complex, monoplex networks become an oversimplification of the corresponding systems [B+14]. This imposes a need to go beyond traditional networks into a richer framework capable of hosting objects and relations of different scales [TBK+16], called Multilayered Network. These complex networks have contributed in many contexts and fields [k+14], although they have been rarely exploited in the investigation of biological networks, where they are very applicable. [DPA+15] In order to fill this gap, we aim to implement a multilayer framework that can be applicable in various domains, especially in the field of pathway modelling. Our idea is to integrate pathways and their related knowledge into a multilayer model, where each layer represents one of their elements. The model offers a feature we call "Selective Inclusion of Knowledge", as well as a collection of related knowledge into a single graph, like diseases and drugs. In this poster, we give an overview of the various models of multilayered networks, then we describe the model we are building, and the workflow of implementing it into an R package as well as the future plan.

References:
[k+14] Kivelä et al., Multilayer Networks. Journal of Complex Networks (2014) 2, 203271
[DPA+15] De Domenico, Porter, and Arenas, MuxViz: a tool for multilayer analysis and visualization of networks. Journal of Complex Networks (2015) 3, 159–176
[B+14] Boccaletti et al., The Structure and Dynamics of Multilayer Networks. Physics Reports 544, 1 (2014)
[TBK+16] Traxl, Boers, and Kurths, Deep Graphs - a General Framework to Represent and Analyze Heterogeneous Complex Systems across Scales. Chaos 26, 065303 (2016)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Machine learning driven time series clustering of heterogeneous single-cell transcription

Sofya Lipnitskaya, Stephan Baumgaertner, Christoph Fritzsch and Stefan Legewie
Institute of Molecular Biology gGmbH
S.Lipnitskaya@imb-mainz.de

Background: During the last years, stochasticity in gene expression has been extensively studied and was found to play a crucial role in cellular decision making. However, currently, there is no standardized procedure to classify single-cell mRNA expression time courses into subpopulations with similar characteristics of cellular variability. Estrogen is a steroid hormone, which regulates cell proliferation in breast cancers. Using estrogen-dependent transcription as a model system, this study aims to develop an effective computational approach, which is able to identify functionally related groups of cells based on their transcriptional patterns and analyze the effect of heterogeneity on intra-tumor variability at single-cell resolution. Methods: Single-cell nascent mRNA time-series data were obtained for the estrogen-responsive GREB1 gene in MCF-7 breast cancer cells using live-cell imaging. To define subpopulations of cells that reacted differently to the same stimulus, a clustering workflow was developed which includes extraction of statistical features from the time series, application of nonlinear dimensionality reduction (multidimensional scaling, isomap) and clustering (hierarchical, gaussian mixture model) algorithms. This workflow was optimized and its performance assessed using an artificial dataset, in which subpopulations were simulated using a stochastic gene expression model developed in our group. The final accuracy evaluation was performed using conditional entropy-based external cluster metrics which allows to determine whether cluster assignments of the samples satisfy predefined labels. Results: It was found that MCF7 cells are inherently variable in GREB1 expression levels over a wide range of estrogen concentrations, with cells reacting very differently even if they are subjected to the same stimulus. Hidden subpopulations could be identified in artificial and real datasets, each group exhibiting specific characteristics of stochastic transcriptional bursting. Conclusions: The suggested clustering approach is able to analyse heterogeneous time-series data, and provides a basis for identifying causes and consequences of cell-to-cell variability in cancer cell subpopulations.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Computational prediction of complexome profile maps based on protein complex assembly models

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

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

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

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

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

Complexome profiling [HBS+12] is a modern technique to study the composition and dynamics of native macromolecular complexes and supercomplexes [SKH+16]. Recently, a dynamic complexome profiling (CP) approach has been applied to elucidate the assembly of human mitochondrial complex I [GCBK+17]. This has been achieved by the incorporation of multiple CP snapshots taken at different time points during the assembly of human mitochondrial complex I. The CP maps of the different time points have been compared manually to reconstruct the order in which the proteins of complex I assemble into subunits, the functional complex and supercomplex structures. This manual comparison of the CP datasets is a time-consuming process and would benefit greatly from computational assistance.

We are developing a software suit for the automated prediction of assembly models for protein complexes based on CP data. From this suit we introduce our software for the validation of assembly models. The software creates an expected CP map from a given assembly model. The resulting expected CP map can be compared to the experimental CP maps that were used for the prediction of the assembly model. The functionality of the software is demonstrated with a test case.

References

[GCBK+17] Sergio Guerrero-Castillo, Fabian Baertling, Daniel Kownatzki, Hans J. Wessels, Susanne Arnold, Ulrich Brandt, and Leo Nijtmans. The Assembly Pathway of Mitochondrial Respiratory Chain Complex I. *Cell Metabolism*, 25(1):128 – 139, 2017.

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

[SKH+16] Valentina Strecker, Zibirnisa Kadeer, Juliana Heidler, Cristina-Maria Cruciat, Heike Angerer, Heiko Giese, Kathy Pfeiffer, Rosemary A. Stuart, and Ilka Wittig. Supercomplex-associated Cox26 protein binds to cytochrome *c* oxidase. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1863(7, Part A):1643 – 1652, 2016.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Jump-starting Tumorigenesis ?
Transposition Activity in Glioblastoma

Konrad Grützmann[2,6], Falk Zakrzewski[2,3,6], Alexander Krüger[2,6], Mathias Lesche[4], Evelin Schröck[1,2], Barbara Klink[1,2], Guido Reifenberger[5,6], Daniela Aust[2,3]

1 Department for Clinical Genetics, University Hospital Carl Gustav Carus Dresden, Faculty of Medicine Carl Gustav Carus, Techincal University Dresden, Germany

2 Core Unit for Molecular Tumor Diagnostics (CMTD) at the National Center for Tumor Diseases (NCT) partner site Dresden, Germany

3 Institute for Pathology, University Hospital Carl Gustav Carus Dresden, Germany

4 Deep Sequencing Group SFB 655, Biotechnology Center, Technische Universität Dresden, Germany

5 Department of Neuropathology, Heinrich Heine University Düsseldorf, German Cancer Consortium, Essen/Düsseldorf, Germany

6 German Cancer Research Center (DKFZ), Heidelberg, Germany

correspondence: konrad.gruetzmann@uniklinikum-dresden.de

Glioblastoma multiforme (GBM) is considered the most common aggressive primary brain tumor in adults because it rapidly spreads into neighboring brain tissue. Up to now there is no targeted therapy, leaving resection, radiotherapy and chemotherapy as the standard treatment. GBMs return after treatment resulting in a median patient survival of about 18 months [Mea17]. Despite therapeutic advances, heterogeneity, tumor evolution and drug-resistance pose an enduring challenge. Thus, research for prognostic markers and therapeutic targets has intensified in recent years.

About half of the human DNA is comprised of repeated sequence derived from transposable elements (TEs). Many of these have profound effect on the genetic buildup [Bur17]. Especially LINE-1 mediated retrotransposition is accepted as a major contributor to structural variation in humans, e.g. via disruption of coding sequence, aberrant expression of newly linked genes, and chromosomal rearrangements [Bur17]. LINE-1-mediated retrotransposition likely accounts for 1 in every 250 pathogenic human mutations [Kea17]. Changed expression and mobility of LINE-1 is typical for cancer, and can drive mutations in tumorigenesis [Bur17].

Only few aspects of TEs in GBM have been investigated so far. Retrotransposition is rare in GBM [Bur17]. However, *in vitro* transposon mutagenesis showed that neural stem cells can be transformed into precursors of glioma-initiating cells by altering GBM-associated genes [Kea12]. TEs may also be pieces in the puzzle of GBM evolution and drug-resistance. Hence GBM subtype-specific TEs could serve as prognostic markers and therapeutic targets, e.g. in curative epigenetic TE suppression.

We here show the analysis of transcribed TEs from RNA-Seq of 4 healthy control and 16 GBM samples, including 4 different GBM subtypes. Cluster analysis shows that GBM subtypes have specific TE expression patterns. We present differential TE expression analysis and the relation to clinical parameters of patients, e.g. survival time. To approach the role of TEs in mutagenesis, we reconstruct structural rearrangements from chimeric RNA-Seq reads involving TEs and non-TE genomic loci.

References

[Bur17] Kathleen H Burns. Transposable elements in cancer. *Nature reviews. Cancer*, 17:415–424, July 2017.

[Kea12] Hideto Koso et al. Transposon mutagenesis identifies genes that transform neural stem cells into glioma-initiating cells. *Proceedings of the National Academy of Sciences of the United States of America*, 109:E2998–E3007, October 2012.

[Kea17] Haig H Kazazian et al. Mobile DNA in Health and Disease. *The New England journal of medicine*, 377:361–370, July 2017.

[Mea17] Ana Miranda et al. Breaching barriers in glioblastoma. Part I: Molecular pathways and novel treatment approaches. *International journal of pharmaceutics*, July 2017.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

MetaGraph: De Bruijn graph based data structures and algorithms for comparative and metagenomics

Andreas Andrusch, Michael Schwabe, Simon H. Tausch, Piotr W. Dabrowski, Bernhard Y. Renard, Andreas Nitsche
Robert Koch Institute, Berlin, Germany
andruscha@rki.de

Due to the ever growing amount of NGS data generated, efficient data structures for their storage and analysis are becoming increasingly crucial. Here we present MetaGraph, a novel approach addressing both requirements in the context of metagenomic data analysis. MetaGraph uses a de Bruijn graph based data structure for reference sequence storage augmented with sequence metadata, including their taxonomic lineage. One of MetaGraphs main applications is the taxonomic binning of unknown sequences, for example reads from metagenomic NGS datasets, in order to assess sample constituents. Additionally, it enables classifications and comparisons based on user selectable clades, stepping away from single references towards pan-genome references. Outside of the field of metagenomics, it enables the researcher to perform a wide array of analyses important for comparative genomics. This includes sequence comparisons of references against reads or other references in order to find shared sequence stretches or unique subsequences. In the same fashion it allows the analysis of pan-genomes. Using the graph structures presented here, MetaGraph avoids redundant computations by fully exploiting similarities between sequences. Due to its flexibly extensible sequence metadata it can be adapted to a multitude of sequence analysis contexts. These features are realized using a highly performant and scalable data structure, which utilizes sequence redundancies to amortize space requirements. It performs comparably or better than published tools with similar functionality in both speed and scalability. Further improvements are planned regarding the out-of-process storage of MetaGraphs data structures using database back-ends and increases in sensitivity by working with spaced k-mers.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

mitoBench & mitoDB: Novel interactive methods for population genetics on mitochondrial DNA

Judith Neukamm, Alexander Peltzer, Wolfgang Haak and Kay Nieselt
Center for Bioinformatics (ZBIT), University of Tübingen, Germany.
judith.neukamm@uni-tuebingen.de

Despite the availability of modern next generation sequencing technologies and therefore nuclear human genomes, for many applications and population genetics studies, the sequencing and analysis of mitochondrial DNA (mtDNA) is still common. Especially in the research field of ancient DNA, mtDNA is often the only proxy available to study extinct populations and their relationship with modern populations.

A plethora of methods for the analysis of mtDNA exist that address questions in population genetics, phylogeny and others. However, these tools typically rely on different file formats and often require manual interaction with the data for downstream analysis. Ultimately, these steps can be cumbersome, especially for non-bioinformaticians with an increased risk of errors during the analysis.

Our ultimate aim is to provide a central reference database for population genetics studies on complete mitogenomes that can be easily accessed both via a web interface and the accompanying mitoBench application, enabling users to perform typical analysis procedures much faster and more conveniently than before.

Here, we present mitoBench and mitoDB. MitoBench is a workbench to interactively analyse and visualize mitochondrial genomes with a focus on population genetics. The graphical user interface is kept simple, to accommodate even users without further prior knowledge on computational methods. Furthermore, it shows additional information such as metadata and statistics. Currently, mitoBench offers automatic file conversion tools to connect the workbench with common analysis methods such as BEAST, Arlequin and others. It also provides basic downstream analysis methods to investigate correlations between populations, such as principal component and FST analysis.

MitoDB aims at providing a large reference panel of modern and ancient mitogenomes for population genetics. The current prototype provides a basis of several thousand complete mitogenomes from the 1000 Genomes project and is constantly extended. The whole system is implemented to be a free web-service that provides interactive and exploratory access to the database itself.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

gemPlot: the first implementation of the 3-dimensional boxplot, as a tool for outlier detection in high-throughput expression data

Jochen Kruppa and Klaus Jung
Institute for Animal Breeding and Genetics, TiHo Hannover
Jochen.Kruppa@tiho-hannover.de

Molecular high-throughput expression data (e.g., microarray or RNA-seq) that represents multiple patient groups often contains outliers which makes the analysis little robust. I.e., results are very sensitive to the addition or removal of a single individuals. The identication of outlying or extreme observations is an very important step of quality control before doing further data analysis. Outlier detection methods for univariate data are however not applicable, since high-dimensional expression data includes usually thousands of features observed in small samples. Existing methods are mainly based on visual inspection of hierarchical cluster trees or principal component plots. Pure visual approaches depend, however, on the individual judgement of the analyst and are hard to automate. Furthermore, currently available methods for automated outlier detection are only applicable to data of a single patient group.

We present the gemplot [KJ17] as the 3-dimensional extension of the 1-dimensional boxplot and the 2-dimensional bagplot [RRW99] as a tool for automated outlier detection in molecular high-throughput data. Bagplots or gemplots can be applied, separately to the data of each study group, after dimension reduction by principal component analysis. Bagplots and gemplots surround the regular observations with convex hulls and observations outside these hulls are regarded as outliers. The convex hulls are determined separately for the observations of each experimental group while the observations of all groups can be displayed in the same subspace of principal components. The applicability of our method to multigroup data is a clear advantage over other available methods. We provide an implementation of the gemplot in the R-package 'gemPlot' available from GitHub (https://github.com/jkruppa/gemPlot).

References:

[KJ17] Jochen Kruppa and Klaus Jung. Automated multigroup outlier identication in molecular high-throughput data using bagplots and gemplots. BMC Bioinformatics, 18(1), may 2017.

[RRW99] P J Rousseeuw, I Ruts, and Tukey J W. The Bagplot: A Bivariate Boxplot. Am Stat, 53(4), 1999.gemPlot)

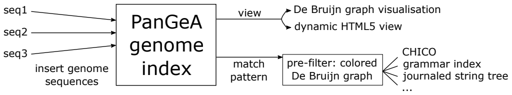
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

PanGeA: Pan-Genome Annotation
Indexing Annotated Human Genome Collections

Andre Brehme, Sven Brümmer, Jonas Charfreitag, Jonas Ellert, Jannik Junghänel, Dominik Köppl,
Christopher Osthues, Sven Rahmann, Dennis Rohde, Julian Sauer, Jonas Schmidt, Lars Schäpers,
Uriel Elias Wiebelitz, Jens Zentgraf
Chair for Algorithm Engineering, Technical University of Dortmund
pgpangea.cs@lists.tu-dortmund.de

Novel high-throughput sequencing methods make it possible to create huge sets of genomes. A genome of these sets is nowadays often found to have annotations of their gene sequences. By the study and analysis of the similarities and differences of an annotated area among the genomes of individuals, opportunities for personalized genome-based medicine arise.

The genome of a human individual consists of roughly 3 billion base pairs that can be stored plainly in 1 GB of space, which is possible because each base pair takes 2 bits. To provide a tool for genome-based medicine, the genomes of all patients have to be stored and maintained in a sophisticated index that supports answering queries efficiently, while compressing the input genome sequences. We highlight the latter point as crucial, for instance because just storing the population of a middle-sized city like Dortmund already costs 500,000 GB of space. Our idea to compress the input is to exploit the circumstance that two individuals of the same species share around 99% of the genetic information [FCS06].



This high inter-similarity of the genomes motivated us to devise a novel index data structure that explicitly takes advantage of the facts that (a) the input consists of a collection of DNA sequences that are highly similarly to each other, and that (b) the annotations can help to cluster together common gene sequences. To make this all possible, we have implemented different indexes and compression techniques so far, like a colored De Bruijn [BBB+17] graph as a pre-filter, the JST [RWR14] and a tailored-version of CHICO in [Val16] and relative Lempel-Ziv [KPZ11].

As an outlook, we currently work on combining one of the aforementioned indexes with compression and on implementing other approaches like grammar-based self-indexes [CN12]. In addition, we plan to use the additional information of the annotations for annotation based queries. For example it should be possible to request in what area of the sequences of the indexed pangenome a specific variation of a gene is contained.

References

[BBB+17] Keith Belk, Christina Boucher, Alexander Bowe, Travis Gagie, Paul Morley, Martin D Muggli, Noelle R Noyes, Simon J Puglisi, and Rober Raymond. Succinct Colored de Bruijn Graphs. *Bioinformatics*, to appear, 2017.

[CN12] Francisco Claude and Gonzalo Navarro. Improved Grammar-Based Compressed Indexes. In *Proc. SPIRE*, volume 7608 of *LNCS*, pages 180–192. Springer, 2012.

[FCS06] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. Structural variation in the human genome. *Nature Reviews. Genetics*, 7(2):85–97, 2006.

[KPZ11] Shanika Kuruppu, Simon J. Puglisi, and Justin Zobel. Relative Lempel-Ziv Compression of Genomes for Large-scale Storage and Retrieval. In *Proc. SPIRE*, volume 7024 of *LNCS*, pages 201–206. Springer, 2011.

[RWR14] René Rahn, David Weese, and Knut Reinert. Journaled string treea scalable data structure for analyzing thousands of similar genomes on your laptop. *Bioinformatics*, 30(24):3499–3505, 2014.

[Val16] Daniel Valenzuela. CHICO: A Compressed Hybrid Index for Repetitive Collections. In *Proc. SEA*, volume 9685 of *LNCS*, pages 326–338. Springer, 2016.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

megSAP – diagnostic analysis pipeline for NGS data

Christopher Schroeder, Franz Hilke, Alexandra Dhring, Ulrike Faust, Tobias Haack and Marc Sturm
Institute of Medical Genetics and Applied Genomics
christopher.schroeder@med.uni-tuebingen.de

Next-Generation-Sequencing is widely used in clinical diagnostics and translational research. The amount of data generated increases steadily and sequencing costs continue to fall. At the same time, the need for bioinformatic support is growing and specialized software is needed for automated, reproducible data analysis in a high-throughput setting. Thus, we developed megSAP, a free-to-use open-source analysis pipeline for diagnostic or research applications. megSAP offers quality control on several analysis levels and integrates free-to-use databases like 1000 Genomes, ExAC, Kaviar and ClinVar for annotation of variants. Optionally, commercial databases (OMIM, HGMD and COSMIC) can be used if a license is available. Due to the comprehensive annotation, the variant lists (produced in VCF and TSV format) can be easily filtered to identify variants of interest. megSAP is updated on a regular basis (both tools and annotation databases). Each release is validated using the GiaB NA12878 gold-standard dataset, inter-laboratory comparisons and EMQN test schemes. Currently, megSAP can analyze single-sample as well as tumor-normal pair NGS data from whole-genome sequencing, whole-exome sequencing and panel sequencing (both shotgun and amplicon-based data). Several other pipelines (RNA-Seq, trio sequencing, and molecular barcodes) are supported as well. To facilitate the installation of megSAP and thereby improve usability, we are working on a first containerized release using Docker.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Random access to sequence graphs stored in large GFA files

Giorgio Gonnella and Stefan Kurtz
Universität Hamburg, MIN-Fakultät, ZBH - Center for Bioinformatics, Hamburg, Germany
gonnella@zbh.uni-hamburg.de

The GFA (Graphical fragment assembly) formats GFA1 and GFA2 are emerging formats for the representation of sequence graphs, including assembly and variation graphs.

The graph structure of a GFA file is recorded in nodes (segments, representing sequences) connected by different kind of arcs (links, containments, edges and gaps). In GFA files no particular order of the lines is required and lines defining arcs contain two references to segments. For these reason, traversing the graph usually requires to store the contents of the entire GFA file in memory. For large GFA files this becomes infeasible due to limitations of the memory.

We present a method for traversing a GFA graph, without reading the entire GFA file in memory. In particular, we implemented a software tool, which, in a preliminary phase employs an external sorting method to handle GFA files of unlimited size. Thereby, arcs referring to the segment as their first segment reference are sorted immediately after the segment line. The tool also outputs an index containing, for each segment name, the position of the segment line in the sorted GFA file, and a list of positions of the arcs referring to the segment as their second segment position.

As an example application, we implemented a tool, which extracts a subgraph from a GFA file, from a specified segment and traverses the arcs in breadth-first fashion, until a specified depth is reached. Thereby, only the index file must be kept in memory. The advantage of using the index in terms of memory requirement is particularly significant if the GFA file contains long segment names, or data other than the topology information, such as metadata, sequences, alignments, or assignments of reads to contigs.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Rapid Domain Annotation

Carsten Kemena, Elias Dohmen, Erich Bornberg-Bauer
Institute for Evolution and Biodiversity, WWU Münster
c.kemena@uni-muenster.de

Domains are building blocks of proteins and play an important role in the evolution of new proteins. The ability to rearrange domains into different combinations allows to create new functionality without the need to invent new proteins from scratch. Therefore, with a few thousand domains, an enormous number of functionally diverse proteins can be created with relatively simple genetic operations such as duplications, fusions of domain arrangements and terminal losses e.g. by shortening a reading frame.

The most common way to represent a domain is by Hidden Markov Models (HMMs), as for example in the Pfam [FCE+16] database. A domain HMM can be used to annotate sequences with domain instances using a threshold. The high accuracy of HMMs is computationally relatively expensive, especially when annotating large sequence sets such as genomes or large transcriptomes with many different domains.

Considering the huge amounts of data from NGS projects this is becoming a daunting challenge. One way to solve this problem is to represent domains as a collection of words in a database against which can be matched. As no alignment is needed but only a simple match finding algorithm the approach is much faster than using HMMs. An example for such a method was UProC [Mei14] which, however, is not longer maintained and lacked position information. Accordingly, domain rearrangements could not be characterised too well.

We developed RADIANT (RApid DomaIn ANoTation) which also stores words of known domains in a database. RADIANT is also able to determine the order of domain occurrences in a protein which makes it possible to rapidly transform protein sequences in strings of domains for further rapid analysis such as DOGMA [DKBBK16]. Furthermore, memory requirements are smaller than with earlier programmes because our simplified matching algorithm allows to reduce the amount of entries needed to be stored in the domain word database.

We used RADIANT in combination with DOGMA, a program which rapidly estimates proteome and transciptome quality and completeness e.g. from NGS data sets based on domain occurrence. Our results show that proteome quality estimates based on domain annotations with RADIANT are comparable to the corresponding results based on the original PfamScan annotations.

References

[DKBBK16] Elias Dohmen, Lukas P M Kremer, Erich Bornberg-Bauer, and Carsten Kemena. DOGMA: domain-based transcriptome and proteome quality assessment. *Bioinformatics*, May 2016.

[FCE+16] Robert D Finn, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A Salazar, John Tate, and Alex Bateman. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, 44(D1):D27985, January 2016.

[Mei14] Peter Meinicke. UProC: tools for ultra-fast protein domain classification. *Bioinformatics*, December 2014.

Integrating spatial, morphological and textural information for improved cell type differentiation using Raman microscopy

Sascha D. Krauß*[1], Hesham K. Yosef[1], Tatjana Lechtonen[1], Hendrik Jütte[2], Andrea Tannapfel[2],
Heiko U. Käfferlein[3], Thomas Brüning[3], Florian Roghmann[4], Joachim Noldus[4],
Samir F. El-Mashtoly[1], Klaus Gerwert[1], and Axel Mosig[1]

[1]Department of Biophysics, Ruhr-University Bochum, 44780 Bochum, Germany
[2]Bergmannsheil Hospital, Ruhr-University Bochum, 44789 Bochum, Germany
[3]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance,
Institute of the Ruhr-University Bochum (IPA), 44789 Bochum, Germany
[4]Department of Urology, Marien Hospital Herne, Ruhr-University Bochum, 44625 Herne, Germany

Raman microscopy is a well-established tool for distinguishing different cell types in cell biological or cytopathological applications, since it can provide maps that show the specific distribution of biochemical components in the cell, with high lateral and spatial resolution. Currently, established data analysis approaches for differentiating cells of different types mostly rely on conventional chemometrics approaches, which tend to not systematically utilize the advantages provided by Raman microscopic data sets. To address this, we propose two approaches that explicitly exploit the large number of spectra as well as the morphological and textural information that are available in Raman microscopic data sets.

Spatial bagging as our first approach is based on a statistical analysis of majority vote over classification results obtained from individual pixel spectra. Based on the *Condorcet's Jury Theorem* (CJT), this approach raises the accuracy of a relatively weak classifier for individual spectra to nearly perfect accuracy at the level of characterizing whole cells. Our second approach extracts morphological and textural (*morpho-textural*) features from Raman microscopic images to differentiate cell types. While using few wavenumbers of the Raman spectrum only, our results indicate on a quantitative basis that Raman microscopic images carry more morphological and textural information than hematoxylin and eosin (H&E) stained images as the current gold standard in cytopathology. Our two approaches promise improved protocols for the fast acquisition of Raman imaging data, for instance for the morphological analysis of *coherent anti-Stokes Raman spectroscopy* (CARS) microscopic imaging data, or for improving the accuracy of fibre optical probe systems by resampling spectra and utilizing spatial bagging.

KEYWORDS
Raman microscopy, cytopathology, supervised learning, spatial bagging, morphological classification

References

[KYL+17] Sascha D Krauß, Hesham K Yosef, Tatjana Lechtonen, Hendrik Jütte, Andrea Tannapfel, Heiko Udo Käfferlein, Thomas Brüning, Florian Roghmann, Joachim Noldus, Samir F El-Mashtoly, Klaus Gerwert, and Axel Mosig. Integrating spatial, morphological and textural information for improved cell type differentiation using Raman microscopy (under revision, manuscript number CEM-17-0125). *J. Chemometrics*, 2017.

[YKL+17] Hesham K Yosef, Sascha D Krauß, Tatjana Lechtonen, Hendrik Jütte, Andrea Tannapfel, Heiko Udo Käfferlein, Thomas Brüning, Florian Roghmann, Joachim Noldus, Axel Mosig, Samir F El-Mashtoly, and Klaus Gerwert. Non-invasive diagnosis of high-grade urothelial carcinoma in urine by Raman spectral imaging. *Anal. Chem.*, 89(12):6893–6899, 2017.

*sascha.krauss@bph.rub.de

Interactive Pangenome Visualization Using Variant Graphs

Simon Heumos[1,2], Björn Geigle[2], Theodore R Gibbons[2], Jörg Hagmann[2], Sebastian J Schultheiss[2], Verena JW Schünemann[3] and Daniel Huson[1]
[1] *Algorithms in Bioinformatics, University of Tübingen;*
[2] *Computomics GmbH;*
[3] *Department of Archaeological Sciences, University of Tübingen*
agv@computomics.com

The steadily declining cost of sequencing and assembling multiple genomes from a single species has created a growing need for intuitive, visual, comparative exploration of pangenome data. Currently available genome viewers limit representations by fixing one genome as a global reference, against which all others are compared. This can make it difficult to identify and compare subpopulations to which the global reference does not belong. Recent efforts have focused on variant graph representations, in which individual genomes are defined by paths through a graph that encodes genomic subsequences as nodes.

Here, we present the Augmented Genome Graph Visualization (AGV), a webbased visualization server to explore variants of whole populations. Compared to existing genome browsers, its data structure is based on the opensource project variation graph (vg), which efficiently stores genomes using a lossless compression. AGV enables a deep exploration of large genome population datasets, including the intuitive display of large structural variants. This webbased clientserver model performs data storage and intensive computation on the server, while visualizations are performed in the clients web browser. Information can be shared between collaborators without installing any local software. AGV includes liftover of reference annotations and the calculation and visualization of haplotype blocks[WP03], i.e. regions of strong linkage disequilibrium that indicate genome regions of low recombination frequency.

References

[WP03] J. D. Wall and J. K. Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetic*, 4(8):587–597, 2003.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Analysis and Comparison of the Ubiquitinome and Phosphoproteome of Salmonella-infected Cells

Jens Rieser
Goethe-Universitt Frankfurt
jens.rieser@bioinformatik.uni-frankfurt.de

Salmonella Typhimurium provokes gastroenteritis and typhoid fever and causes many thousands death every year. A better understanding of the host-pathogen interaction can lead to a better medical therapy. The post-translational modification of proteins after an infection is one aspect of interest. Ubiquitination and phosphorylation are two reversible possibilities for a quick answer of the cell to environmental changes. Phosphorylation for example 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 the ubiquitination and phosphorylation of proteins during a Salmonella infection we analyzed the data of two different datasets. The changes of the expression level of phosphorylated proteins in Salmonella-infected and uninfected cells has been investigated by Rogers et al. [2] and the changes in ubiquitinated proteins by Fiskin et al. [3]. We used the proteins of both datasets to search for protein-protein interactions in three databases: STRING, IntAct and BioGRID. The received interactions were used to create protein-protein interaction networks for nucleus with 1,646 proteins and 17,875 interactions, membrane with 1,255 proteins and 11,137 interactions and cytoplasm with 1,704 proteins and 18,978 interactions. The amount of proteins with data from both, ubiquitination and phosphorylation, in the nucleus network was 270 (16.4%) proteins, in the membrane network 219 (17.4%) proteins and in the cytoplasm network 230 (13.5%) proteins. Every network was clustered and analyzed according to GO enrichments. Interesting clusters were found for example in cytoplasm network with 48 proteins and the GO term “NIK/NF-B signaling” and a P-value of $1.39E-28$. To take the ubiquitination and phosphorylation back into account, the corresponding measurements were mapped to each protein for a direct comparison in the network.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Bringing Pathway Knowledge to Systems Medicine Approaches

Florian Auer, Tim Beißbarth and Frank Kramer
Department of Medical Statistics, University Medical Center Göttingen
florian.auer@med.uni-goettingen.de

In modern Systems Medicine approaches the aim is to look at increasingly complex interactions of complete signaling pathways in order to get a more holistic view for individualized treatment decisions. Individualized treatment decisions and newly developed specialized drugs warrant the need to broaden the focus in individualized medicine from singular biomarkers to pathways.

On the other hand pathway databases offer vast amounts of knowledge on biological networks, freely available and encoded in semi-structured formats[BCS06, SAK$^+$09]. The efficient re-use of pathway knowledge and its integration into bioinformatic analyses enables new insights for researchers in systems medicine.

However, the vast amount of published data on molecular interactions makes it increasingly challenging for life science researchers to find and extract the most relevant information. Currently, the tools to use this information and integrate it in a clinical context are still lacking.

Our idea is to compose an analysis pipeline in order to enable patient-specific systems medicine analyses in a university hospital setting. Our poster will present a workflow for visualizing pathway information and integrating omics data within an interactive online application, utilizing state of the art technology[FLH$^+$16, R C14, KBK$^+$13, FBBL15] and well-established standard data models[DCP$^+$10, HFS$^+$03, PCW$^+$15].

References

[BCS06] Gary D. Bader, Michael P. Cary, and Chris Sander. Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(Database issue):D504–506, January 2006.

[DCP$^+$10] Emek Demir, Michael P Cary, Suzanne Paley, et al. The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942, September 2010.

[FBBL15] Silvia Frias, Kenneth Bryan, Fiona S. L. Brinkman, and David J. Lynn. CerebralWeb: a Cytoscape.js plug-in to visualize networks stratified by subcellular localization. *Database*, 2015:bav041, January 2015.

[FLH$^+$16] Max Franz, Christian T. Lopes, Gerardo Huck, et al. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, 32(2):309–311, 2016.

[HFS$^+$03] M. Hucka, A. Finney, H. M. Sauro, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, March 2003.

[KBK$^+$13] F. Kramer, M. Bayerlova, F. Klemm, A. Bleckmann, and T. Beissbarth. rBiopaxParser–an R package to parse, modify and visualize BioPAX data. *Bioinformatics*, 29(4):520–522, February 2013.

[PCW$^+$15] Dexter Pratt, Jing Chen, David Welker, et al. NDEx, the Network Data Exchange. *Cell Systems*, 1(4):302–305, October 2015.

[R C14] R Core Team. R: A Language and Environment for Statistical Computing. 2014.

[SAK$^+$09] Carl F. Schaefer, Kira Anthony, Shiva Krupa, et al. PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(Database issue):D674–D679, January 2009.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Manatee invariants for the analysis of bipartite graphs in systems biology

Leonie Amstein, Jennifer Scheidel, Jörg Ackermann, and Ina Koch
Molecular Bioinformatics Group, Institute of Computer Science, Goethe-University Frankfurt am Main, Germany

Biological systems like metabolic or signaling pathways form highly intertwined networks. Models allow for simulation and analysis to understand system-wide cellular behavior. Some important system's properties can be only examined on a systems level, such as robustness and vulnerability.

Petri nets are directed, labeled, bipartite graphs. Petri nets are suitable to describe and analyze biological processes in detail. The Petri net formalism provides profound methods to analyze the dynamic behavior based on transition invariants without the need for extensive kinetic parameters [KRS11]. Transition invariants correspond to the concept of elementary modes [SFD00] and reveal processes in the model that occur under steady-state conditions. Cycles in the topology of a model have an important impact on the biological interpretation of the results of the transition invariant analysis. Cycles disrupt the capturing of signal flows such that the resulting transition invariants capture isolated regulatory processes.

In order to analyze all possible signal flows, we extend the concept of transition invariants. We adapt the concept of feasible transition invariants [SHK06] and define Manatee invariants [AAS$^+$17] to combine transition invariants in the sense that they represent interrelated processes like signal flows. We determine the cycles, which cause the disruption as internal place invariants in the subnetwork of a transition invariant. Manatee invariants minimize the number of place invariants in subnetworks of transition invariants by linear combination of interrelated transition invariants. The resulting Manatee invariant is a transition invariant, which is not minimal but has a subnetwork that is free of place invariants.

The concept of Manatee invariants employs the properties of the bipartite graph to treat cycles in the topology of models and determine interrelated biological processes such as signal flows. Especially in models of signaling pathways, the determination of the combinatorial diversity of signal flows is elementary for a rigorous model analysis such as for crosstalks, feedback loops, and *in silico* knockouts [SAA$^+$16]. The mathematical framework of Manatee invariants advances analysis of robustness and vulnerability of models in systems biology.

References

[AAS$^+$17] Leonie Amstein, Jörg Ackermann, Jennifer Scheidel, Simone Fulda, Ivan Dikic, and Ina Koch. Manatee invariants reveal functional pathways in signaling networks. *BMC Systems Biology*, 2017.

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

[SAA$^+$16] Jennifer Scheidel, Leonie Amstein, Jörg Ackermann, Ivan Dikic, and Ina Koch. *In Silico* Knockout Studies of Xenophagic Capturing of *Salmonella*. *PLoS Computational Biology*, 2016.

[SFD00] Stefan Schuster, David Fell, and Thomas Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 2000.

[SHK06] Andrea Sackmann, Monika Heiner, and Ina Koch. Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, 2006.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Building an open, collaborative, online infrastructure for bioinformatics training

Berenice Batut[1], Galaxy Training Network[2], Dave Clements[3], Bjoern Gruening[1]
[1] *University of Freiburg, Germany*
[2] *https://galaxyproject.org/teach/gtn/*
[3] *Johns Hopkins University, United States of America*
berenice.batut@gmail.com

With the advent of high-throughput platforms, life science data analysis is tightly linked to the use of bioinformatics tools, resources, and high-performance computing. However, the scientists who generate the data often do not have the knowledge required to be fully conversant with such analyses. To involve them in their own data analysis, these scientists must acquire bioinformatics vocabulary and skills through training.

Data analysis training is particularly challenging without a computational background. The Galaxy framework is addressing this problem by offering a web-based, intuitive and accessible user interface to numerous bioinformatics tools.

Recently, the Galaxy Training Network (GTN) set up a new open, collaborative, online model for delivering high-quality bioinformatics training material: http://training.galaxyproject.org.

Each of the current 13 topics provides tutorials with hands-on, slides and interactive tours. Tours are a new way to go through an entire analysis, step by step inside Galaxy in an interactive and explorative way. All material is openly reviewed, and iteratively developed in one central repository by almost 50 contributors. Content is written in Markdown and, similarly to Software/Data Carpentry, the model separates presentation from content. In addition, the technological infrastructure needed to teach is described with a list of needed tools, annotation of public Galaxy instances and Docker images for each topic. The data are also stored in Zenodo and citable via DOI.

All materials are annotated by a rich set of metadata (time and resource estimations) and automatically propagated to ELIXIR's TeSS portal. This approach creates tutorials that are accessible, easy to find and (re)use (FAIR) by individuals and by trainers for workshops.

With this community effort, the GTN offers an open, collaborative, FAIR and up-to-date infrastructure for delivering high-quality bioinformatics training for scientists.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Cell orientation in tissue images of diffuse large B-cell lymphoma

Sonja Scharf[1,2], Tobias Bergmann[1], Jörg Ackermann[1], Martin-Leo Hansmann[2] and Ina Koch[1]

[1] *Molecular Bioinformatics, Institute of Computer Science, Goethe-University, Frankfurt am Main*
[2] *Dr. Senckenberg Institute of Pathology, Goethe-University, Frankfurt am Main*
sonja.scharf@med.uni-frankfurt.de

Motivation: In the lymph node, there are different kinds of cells, for example, B-cells and T-cells. These cells interact to trigger an immune response. To enable frequent interactions, cell dynamics is essential. In several videos, such cell movement have been demonstrated [BEK+06], but the major part of information on the lymph node is based on static images which are produced in the daily work of pathologists. Based on static images, we explored, whether the orientation of neighbored cells was correlated. Since a cell deforms to enable its movement in the spatially restricted environment of a lymph node [BEK+06], a correlation of orientation would indicate a collective cell movement. We are interested in the cancer-affected tissue. We analyzed lymph node images of patients diagnosed with Diffuse Large B-Cell Lymphoma (DLBCL), which is the most common lymphoid malignancy (31%) of all non-Hodgkin lymphoma [MFA+13, HED+14].

Methods: We analyzed 819 images of 273 medical cases of DLBCL. For each medical case, three sections from a whole slide image were selected by an expert pathologist. The tissues were stained with hematoxylin and eosin (H&E) to observe the nuclei. An imaging analysis pipeline identified the location of cell nuclei in the images and computed the orientation of the major axis of each profile. For this we applied macros of Fiji [SACF+12] and produced an image overlay of the H&E-stained tissue with a color-coded representation of orientation of the nuclei. We developed a statistical method to evaluate the significance of the orientation observed in the images. So, we classified the images to either a class of 'non-random' images and a second class of images with inconspicuous average orientation.

Results: The imaging pipeline identified about 544,000 cell nuclei in 819 images of DLBCL. The number of cell nuclei varied between 73 and 1568 per image. For a fraction of images, the averaged angle of orientation differed significantly from the value 90 degrees. Note that, an averaged angle of 90 degrees would be expected for randomly orientated cell nuclei. For separating the interesting images, we used the significance levels of 5%, 1%, and 0.1%, the fraction of 'non-random' images were 61%, 51%, and 39%, respectively. The statistical significance of the classification was highest if we chose a significance level of 1% for the orientation. Consequently, the deviation of the mean orientation in an image is a characteristic feature of a medical case and may serve to classify DLBCL into subgroups, together with other features.

References

[BEK+06] Marc Bajénoff, Jackson G. Egen, Lily Y. Koo, Jean Pierre Laugier, Frédéric Brau, Nicolas Glaichenhaus, and Ronald N. Germain. Stromal Cell Networks Regulate Lymphocyte Entry, Migration, and Territoriality in Lymph Nodes. *Immunity*, 25(6):989–1001, 2006.

[HED+14] Sylvia Hartmann, Mine Eray, Claudia Döring, Tuula Lehtinen, Uta Brunnberg, Paula Kujala, Martine Vornanen, and Martin-Leo Hansmann. Diffuse large B cell lymphoma derived from nodular lymphocyte predominant Hodgkin lymphoma presents with variable histopathology. *BMC Cancer*, 14(1):332, May 2014.

[MFA+13] Maurizio Martelli, Andrés J.M. Ferreri, Claudio Agostinelli, Alice Di Rocco, Michael Pfreundschuh, and Stefano A. Pileri. Diffuse large B-cell lymphoma. *Critical Reviews in Oncology/Hematology*, 87(2):146–171, 2013.

[SACF+12] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Prediction of long-range vRNA-vRNA interaction networks between influenza A vRNPs

Daniel Desirò[1,3], Markus Fricke[1,3], Hardin Bolte[2], Martin Schwemmle[2,3], Manja Marz[1,3]
[1] *RNA Bioinformatics and High Throughput Analysis, Friedrich Schiller University Jena, Germany*
[2] *Institute of Virology, Albert Ludwig University of Freiburg, Germany*
[3] *European Virus Bioinformatics Center, Jena, Germany*
daniel.desiro@uni-jena.de

In recent years there have been many studies on the packaging of the eight influenza A viral ribonucleoproteins (vRNPs) into forming the viral genome in the shape of a '7+1' arrangement. While there have been many theories about the underlying mechanisms [FME+12], the explicit packaging process is still unknown, especially when considering the packaging of reassortant viruses between different influenza A virus (IAV) strains. Only a few studies have addressed this issue in detail [LHN+10, GIF+13] but, to our knowledge, there has been no attempted to computationally analyze the involved mechanisms.

Here, we computationally show the existence of some long-range RNA-RNA interaction (LRI) hot spots (HS), that are conserved in five different IAV strains at a high minimum free energy level. To predict these interactions we utilized the recently published LRIscan, which is designed to predict LRIs in full viral genomes [FM16]. Although these HS are present in all tested strains, their interacting mates can vary between different IAV species. Some of these interaction sites also include frame shifts, which could hint to a flexible "packaging network". This could also explain why a specific silent mutation that attenuates one strain has no effect on another [GIF+13].

Our first results show, that LRIscan predicts significant HS located in central regions of the viral RNA (vRNA) segments, contrasting the hypothesis that binding sites are mainly in the 3' and 5' termini of the vRNAs [FME+12]. Three HS are located in the segments coding for the polymerase basic protein 2 (PB2), matrix protein (M) and nonstructural protein (NS) and are of special interest for further studies. We anticipate, that they might play a major role in the packaging process and in reasortant viruses.

With this study, we demonstrate that a computational approach to predict the packaging interactions among vRNPs is feasible. We will further examine the predicted HS and interactions to characterize possible packaging networks between different IAV strains. Ultimately, this will enable us to predict the pathogenic potential and reproductive capacity of recombinant IAVs in the context of packaging.

References

[FM16] Markus Fricke and Manja Marz. Prediction of conserved long-range RNA-RNA interactions in full viral genomes. *Bioinformatics (Oxford, England)*, 32:2928–2935, October 2016.

[FME+12] Emilie Fournier, Vincent Moules, Boris Essere, Jean-Christophe Paillart, Jean-Daniel Sirbat, Catherine Isel, Annie Cavalier, Jean-Paul Rolland, Daniel Thomas, Bruno Lina, and Roland Marquet. A supramolecular assembly formed by influenza A virus genomic RNA segments. *Nucleic Acids Res*, 40:2197–2209, March 2012.

[GIF+13] Cyrille Gavazzi, Catherine Isel, Emilie Fournier, Vincent Moules, Annie Cavalier, Daniel Thomas, Bruno Lina, and Roland Marquet. An in vitro network of intermolecular interactions between viral RNA segments of an avian H5N2 influenza A virus: comparison with a human H3N2 virus. *Nucleic Acids Res*, 41:1241–1254, January 2013.

[LHN+10] Chengjun Li, Masato Hatta, Chairul A Nidom, Yukiko Muramoto, Shinji Watanabe, Gabriele Neumann, and Yoshihiro Kawaoka. Reassortment between avian H5N1 and human H3N2 influenza viruses creates hybrid viruses with substantial virulence. *Proc Natl Acad Sci U S A*, 107:4687–4692, March 2010.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

ndexr - an R package to interface with the Network Data Exchange

Florian Auer, Zaynab Hammoud and Frank Kramer
Department of Medical Statistics, University Medical Center Göttingen, Germany.
florian.auer@med.uni-goettingen.de

Motivation:

Seamless exchange of biological network data enables bioinformatic algorithms to integrate networks as prior knowledge input as well as to document resulting network output. However, the interoperability between pathway databases and various methods and platforms for analysis is currently lacking. NDEx, the Network Data Exchange, is an open-source data commons that facilitates the user-centered sharing and publication of networks of many types and formats.

Results:

Here, we present a software package that allows users to programmatically connect to and interface with NDEx servers from within R. The network repository can be searched and networks can be retrieved and converted into igraph-compatible objects. These networks can be modified and extended within R and uploaded back to the NDEx servers.

Availability:

ndexr is a free and open-source R package, available via GitHub (https://github.com/frankkramer-lab/ndexr) and Bioconductor (http://bioconductor.org/packages/ndexr/)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Analysis of chimeric reads for the detection of RNA-RNA interactions

Richard A. Schäfer and Björn Voß
Institute of Biochemical Engineering University of Stuttgart
bjoern.voss@ibvt.uni-stuttgart.de

The ability of RNA to base-pair with itself and other RNAs is crucial for its function in vivo. While there are reasonable approaches to map RNA secondary structures genome-wide, understanding how different RNAs interact 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 (reviewed in [WMW16]) have been developed. Common to all is that they rely on Psoralen mediated in vivo crosslinking and RNA Proximity Ligation (RPL) [RQS15], which covalently links the interacting RNA strands. Subsequently, the RNA is sequenced using RNA-seq and analyzed with respect to inter- and intramolecular RNA-RNA interactions. The methods that have been used so far implement strict algorithms that lack a sophisticated processing of the reads and tend to miss captured interactions. In this work, we present a general pipeline for the inference of RNA-RNA interactions from raw DDD reads. We applied our pipeline to data from different direct duplex detection and compared our results to the original ones. 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. In order to ensure comparability between the established and future DDD methods there is a need for a standardized pipeline to analyze chimeric reads to infer inter- and intramolecular interactions and to guarantee the reproducibility of the analysis.

References

[RQS15] Vijay Ramani, Ruolan Qiu, and Jay Shendure. High-Throughput Determination of RNA Structure by Proximity Ligation. *Nature Biotechnology*, 33(9):980–984, September 2015.

[WMW16] Chase A. Weidmann, Anthony M. Mustoe, and Kevin M. Weeks. Direct Duplex Detection: An Emerging Tool in the RNA Structure Analysis Toolbox. *Trends in Biochemical Sciences*, 41(9):734–736, September 2016.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

AMPS: A pipeline for screening archaeological remains for pathogen DNA

Ron Hübler, Felix M Key, Christina Warinner, Kirsten Bos, Johannes Krause, Alexander Herbig
Max Planck Institute for Science of Human History, Jena
huebler@shh.mpg.de

Second generation DNA sequencing enables large-scale metagenomic studies. Such analyses are not restricted to present day environmental or clinical samples but can also be applied to molecular data from archaeological remains (ancient DNA) in order to provide insights into the host-bacterial relationships throughout human history. Here we present AMPS (Ancient Metagenomic Pathogen Screening), an automated bacterial pathogen screening pipeline for ancient DNA sequence data that provides straightforward and reproducible information on species identification and authentication of their ancient origin. AMPS consists of a customized version of (1) MALT (Megan ALignment Tool) [1], (2) MaltExtract, a Java tool that evaluates a series of authenticity criteria [2, 5] for a list of target species, and (3) customizable post-processing scripts to identify, filter, and visualize candidate hits from the MaltExtract output.

We evaluated AMPS with DNA sequences obtained from archaeological samples known to be positive for specific pathogens, as well as simulated ancient DNA data [4] from 33 bacterial pathogens of interest spiked into diverse metagenomics backgrounds (soil, archaeological bone, dentine, and dental calculus) [4]. AMPS successfully identified all simulated target pathogens with as few as 50 pathogen DNA sequence reads spiked into 5 million total reads comprising a metagenomic library. In addition, we used these data to assess and compensate for biases resulting from the reference database contents and structure. Finally we compared the performance of AMPS to MIDAS [3] and Kraken [6].

Taken together, AMPS 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.

References

1. Herbig, Alexander, et al. "MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman." BioRxive(preprint) (2016).
2. Key, Felix M., et al. "Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication." Trends in Genetics (2017).
3. Nayfach, Stephen, et al. "An integrated metagenomics pipeline for strain profiling reveals novel patterns of transmission and global biogeography of bacteria." Genome Research 26 (2016): 1612-1625 .
4. Renaud, Gabriel, et al. "gargammel: a sequence simulator for ancient DNA." Bioinformatics 33.4 (2017): 577-579.
5. Warinner, Christina, et al. "A Robust Framework for Microbial Archaeology ." Annual Review of Genomics and Human Genetics (2017).
6. Wood, Derrick E and Steven L Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments." Genome Biology (2014).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Scalable and accessible clustering of ncRNAs based on sequence and secondary structures

Milad Miladi, Eteri Sokhoyan, Torsten Houwaart, Rolf Backofen and Björn Grüning
Bioinformatics Group, University of Freiburg, Germany
miladim@cs.uni-freiburg.de

Background: There are many ncRNAs and regulatory elements whose function is still unknown. RNA sequences with putative but unknown functionality can appear for example in genome-wide screens or experiments such as RNA-seq. Clustering of RNA sequences is currently one of the prevalent approaches for detecting and annotating the function of putative ncRNAs and regulatory elements.

Contribution: Here we present Galaxy-GraphClust, a web-based tool suite for large-scale structural clustering of RNAs based on sequence and structural similarity that is provided via the Galaxy framework. Galaxy-GraphClust is a realization of GraphClust [1] inside Galaxy framework that drastically simplifies the task of clustering large amounts of RNAs by making it possible to:
a) interactively perform the clustering of RNAs via a web interface,
b) support computations on different back-ends ranging from personal computers to large scale computer clusters,
c) integrate the clustering workflow with high-throughput sequencing (HTS) analysis.

Availability: Galaxy-GraphClust is available under: http://github.com/BackofenLab/docker-galaxy-graphclust

[1] Heyne et al., GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinfomatics*, 2012.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Evolution of transcription activator-like effectors in *Xanthomonas oryzae*

Annett Erkes[1], Maik Reschke[2], Jens Boch[2], Jan Grau[1]
[1] *Institute of Computer Science, Martin Luther University Halle–Wittenberg*
[2] *Department of Plant Biotechnology, Leibniz Universität Hannover*
grau@informatik.uni-halle.de

Plant-pathogenic *Xanthomonas* bacteria employ transcription activator-like effectors (TALEs) that bind to the promoter of plant genes and activate their transcription. *Xanthomonas* infections result in a substantial yield loss for many crop plants including rice. The binding domain of TALEs consists of tandem repeats of approximately 34 amino acids. These contain 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.

Different *Xanthomonas* strains possess different repertoires of TALEs, which have likely evolved from a common ancestor TALE. Here, we study the evolution of TALEs from the level of RVDs determining target specificity down to the level of DNA sequence with focus on rice-pathogenic *Xanthomonas oryzae* pv. *oryzae* and *Xanthomonas oryzae* pv. *oryzicola* strains.

We discover that only one to three codon pairs coding for one RVD occur in known TALEs, even though the number of theoretically possible codon pairs is substantially larger. We base our analysis on aligned sequences within classes of significantly similar TALEs from published *Xanthomonas* genomes as generated by AnnoTALE [GRE+15], because TALEs within one class are likely evolutionary related.

We compare the aligned RVDs of class members and find synonymous substitutions only for two RVDs, whereas the remaining substitutions lead to a modification of the RVD. Most frequently, only one nucleotide is substituted between the alternative RVDs among class members.

Although even the flanking sequences of RVDs are highly conserved, the nucleotides at some of the flanking positions show dependencies on the RVD type. We train a classifier for distinguishing RVD types by their flanking region alone and apply it to repeats which show a substitution in the alignment. In the majority of cases, this classifier assigns all such repeats of a common class to the same RVD type, although we observe substitutions in the RVD. Our findings indicate that one way how TALE specificities evolve is by direct base substitutions in RVD codons.

In summary, we find strong indications that TALEs may evolve i) by base substitutions in codon pairs coding for RVDs, ii) by recombination of N-terminal or C-terminal regions of existing TALEs, or iii) by deletion of individual TALE repeats, and we propose putative mechanisms [ERBG17].

We finally study the effect of the presence/absence of TALEs and evolutionary modifications in TALEs on transcriptional activation of putative target genes in rice, and find that even single RVD swaps may lead to considerable differences in activation and that the effect of RVD swaps may depend on the specific target box.

References

[ERBG17] Annett Erkes, Maik Reschke, Jens Boch, and Jan Grau. Evolution of Transcription Activator-Like Effectors in *Xanthomonas oryzae*. *Genome Biology and Evolution*, 9(6):1599–1615, 2017.

[GRE+15] Jan Grau, Maik Reschke, Annett Erkes, Jana Streubel, Richard D. Morgan, Geoffrey G. Wilson, Ralf Koebnik, and Jens Boch. AnnoTALE: bioinformatics tools for identification, annotation, and nomenclature of TALEs from *Xanthomonas* genomic sequences. *Scientific Reports*, 6(21077), 2015.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Network-based identification of gene copy number mutations driving oligodendroglioma development

Josef Gladitz and Michael Seifert
TU Dresden
Josef.Gladitz@tu-dresden.de, michael.seifert@tu-dresden.de

Oligodendrogliomas represent 4-8% of diagnosed primary human brain tumors. All oligodendrogliomas show a characteristic co-deletion of the chromosomal arms 1p and 19q, but little is known about putative driver genes located in these genomic regions. The occurrence of nearly identical co-deletions in individual oligodendrogliomas does not allow to narrow down the exact location of putative driver genes with standard statistical approaches. The aim of our project was to develop a novel network-based approach for the identification of putative driver genes located in the 1p/19q region.

To realize this, we first learned oligodendroglioma-specific gene regulatory networks [1,2] based on gene expression and gene copy number data of 178 oligodendroglioma patients from TCGA [3]. Next, we used network propagation [1,2] to determine impacts of mutation-affected genes (differentially expressed genes within the 1p/19q co-deletion) on the expression of known cancer-relevant pathway genes. Comparisons to impacts obtained under random networks revealed 20 putative driver genes that significantly influence the expression of signaling and metabolic pathways. Several of these genes have already been associated with other types of cancer. Moreover, the two top scoring genes, SLC17A7 and ELTD1, have recently been reported to act as tumor suppressor and oncogene in glioblastomas, a closely related tumor type.

We present the first large-scale computational study to pinpoint novel putative driver genes for oligodendrogliomas. Our results indicate that several putative driver genes are located in the 1p/19q region. Generally, our results suggest that our approach is a valuable tool to identify putative tumor drivers in large chromosomal regions affected by DNA copy number mutations.

[1] M. Seifert, B. Friedrich, and A. Beyer: Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis, Genome Biology, 2016, 17:204.

[2] M. Seifert and A. Beyer: regNet: An R package for network-based propagation of gene expression alterations, https://github.com/seifemi/regNet.

[3] The Cancer Genome Atlas Research Network: Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas, New England Journal of Medicine, 2015, 372(26), 2481-2498.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Modeling and Simulating Protein Hypernetworks

Bianca K. Stöcker[1], Johannes Köster[1,2], Eli Zamir[3] and Sven Rahmann[1]
[1] *Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, Germany*
[2] *Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA*
[3] *Max Planck Institute for Intelligent Systems, Stuttgart, Germany*
Bianca.Stoecker@uni-due.de

Protein interactions are fundamental building blocks of biochemical reaction systems underlying cellular functions. The complexity and functionality of these systems emerge not only from the protein interactions themselves but also from the dependencies between these interactions, e.g., due to allosteric activation, mutual exclusion or steric hindrance. Therefore, formal models for integrating and using information about such dependencies are of high interest.

We present an approach for endowing protein networks with interaction dependencies using propositional logic, thereby obtaining *protein hypernetworks*. The construction of those protein hypernetworks is based on public interaction databases [DY12] and known [CCaS+07] as well as text-mined interaction dependencies [KZR12].

We present an efficient data structure and algorithm to simulate protein complex formation in protein hypernetworks taking the constraints from interaction dependencies into account. Further, we show how to simulate perturbation effects (knockout and overexpression of single or multiple proteins, changes of protein concentrations). The efficiency of the model allows a fast simulation and enables the analysis of many proteins in large networks.

We illustrate the benefits of our model on the human adhesome network, with adjusted simulation parameters to match properties of known human protein complexes [RWL+10]. Through comparing complex formation with known true constraints, randomized constraints and no constraints, we show that the interaction dependencies limit the resulting complex size. Additionally the evaluation includes the influence of perturbations on the complex sizes and on the number of unbound proteins. Further, we analyze how those influences differ regarding the perturbation of varied proteins and we demonstrate that the perturbation of single proteins is propagated through the network and thus has an influence on non-direct interactors of the perturbed protein.

References

[CCaS+07] Arnaud Ceol, Andrew Chatr-aryamontri, Elena Santonico, Roberto Sacco, Luisa Castagnoli, and Gianni Cesareni. DOMINO: a database of domain-peptide interactions. *Nucleic Acids Research*, 35(Database):D557–D560, January 2007.

[DY12] Jishnu Das and Haiyuan Yu. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, 6(1):92, January 2012.

[KZR12] Johannes Köster, Eli Zamir, and Sven Rahmann. Efficiently mining protein interaction dependencies from large text corpora. *Integrative Biology*, 4(7):805, 2012.

[RWL+10] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. CORUM: the comprehensive resource of mammalian protein complexes – 2009. *Nucleic acids research*, 38(suppl 1):D497–D501, 2010.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Building usable full genome variation graphs

Christian Kubica, Felix Bemm, Detlef Weigel
Max Planck Institute for Developmental Biology, Tübingen
christian.kubica@tuebingen.mpg.de

The 1001 Genomes Project generated a polymorphism (SNP) and short structural variant (short SV) map for well over 1000 wild strains (accessions) of *Arabidopsis thaliana*. In addition transcriptome, methylation and phenotypic data for most of the accessions were collected. By utilising long read sequencing technologies to generate de novo assemblies of different diverse *A. thaliana* accessions, we are launching the next phase of this project, in which we will detect and genotype large SVs. First we will shift from a single reference based approach to a multiple genome graph, representing a set of highly diverse *A. thaliana* accessions. Based on this we will detect SVs and subsequently genotype these in the 1001 Genomes Project short read data set.

Most genome graphs are constructed from a multiple whole genome alignment (WGA). Building a WGA however is not trivial and its quality depends on the excess of shared regions to form informative nodes and (super-)bubbles (PNGH) in the graph. The quality of the WGA depends on several factors, with the similarity and the repetitiveness of the aligned sequences being the major ones. The diversity will result in less and smaller alignment blocks, whereas the repetitiveness will lead to multiple alignments. Such a WGA will convert into a highly connected, partially circularized graph that contain almost no usable information as nodes are too short and edges are too abundant to reliably and uniquely anchor superbubbles around interesting structural variants.

Here we propose ways to cope with diverse sequences for graph construction. Our main target is to create a low complexity graph. We alter previous graph construction approaches by focusing on local alignment anchors. The approach reduces the alignment fragmentation by only considering regions near useful alignment anchors (MUMs (DKF+99)/ Minimizers (RHH+04)) and thus prohibits self alignments, which would result in the circularization of the graph. In a second approach we only focus on regions of interest and resolve them to the highest possible resolution and skip non informative parts around them. We further show that in a finished graph, variation can be removed by pruning thus taking information, such as allele frequencies within a population data set, into account. Although our approaches result in loss of information they enable us to generate genome graphs that help to understand variation of SNPs, short and long SVs as well as TEs at an unprecedented resolution when combined with previously generated short read data.

References

[DKF+99] Arthur L Delcher, Simon Kasif, Robert D Fleischmann, Jeremy Peterson, Owen White, and Steven L Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, 1999.

[PNGH] Benedict Paten, Adam M Novak, Erik Garrison, and Glenn Hickey. Superbubbles, Ultrabubbles and Cacti.

[RHH+04] Michael Roberts, Wayne Hayes, Brian R Hunt, Stephen M Mount, and James A Yorke. Reducing storage requirements for biological sequence comparison. *BIOINFORMATICS*, 20(18):3363–3369, 2004.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Automatic sequence refinement via de novo assembly of RNA-Seq data from orthologous species

Julia F. Söllner[1,2], Germán Leparc[1], Matthias Zwick[1], Kay Nieselt[2], Eric Simon[1]
1. Target Discovery Research, Boehringer Ingelheim Pharma GmbH & Co. KG
2. Integrative Transcriptomics, Center for Bioinformatics, University of Tübingen
julia.soellner@boehringer-ingelheim.com

In pharmaceutical research one often requires the precise sequence of a protein to generate recombinant proteins or stable cells lines overexpressing the recombinant protein. These proteins or cell lines can be used to assess the pharmacological effect of a specific drug on its intended target. This information facilitates dose selection for treatment and interpretation of the drugs on-target effectivity and safety. Public databases such as Ensembl, UniProt and RefSeq provide genome-wide sequence information for many higher species. However, this is often based on automatic annotation pipelines and thus incomplete and/or conflicting.

One way to refine such incomplete sequences is based on *de novo* transcriptome assembly using RNA-Seq data. In 2015 Fushan et al.[FTL+15] have published RNA-Seq data for several mammalian species and three tissues (liver, kidney and brain) per species. An in-house RNA-Seq pipeline was applied to derive expression data from the raw data in order to identify highly expressed genes. In addition, we used the fastq files for *de novo* transcriptome assembly with BinPacker[LLC+16] for each of the three tissues in the species of interest.

Based on the expression data and homology information provided by Ensembl, we did a systematic screening for conserved genes with incomplete sequence information in one of the model species and high expression in at least one of the available tissues in the orthologous species.
In order to assess whether a gene's sequence is incomplete we calculated the difference in percent sequence identity of the human protein sequence aligned to its ortholog and the percent identity of the ortholog aligned to the human sequence.

For the resulting genes we aligned the human orthologous protein sequence to the contigs of the target species' respective tissue. From this we obtain a refined protein sequence which was validated via a multiple sequence alignment of the refined sequence and its orthologous sequences.

References

[FTL+15] Alexey A Fushan, Anton A Turanov, Sang-Goo Lee, Eun Bae Kim, Alexei V Lobanov, Sun Hee Yim, Rochelle Buffenstein, Sang-Rae Lee, Kyu-Tae Chang, Hwanseok Rhee, Jong-So Kim, Kap-Seok Yang, and Vadim N Gladyshev. Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, 14(3):352–365, 2015.

[LLC+16] Juntao Liu, Guojun Li, Zheng Chang, Ting Yu, Bingqiang Liu, Rick McMullen, Pengyin Chen, and Xiuzhen Huang. BinPacker: Packing-Based De Novo Transcriptome Assembly from RNA-seq Data. *PLOS Computational Biology*, 12(2):1–15, 02 2016.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Modeling photorespiratory shunts to improve plant crop yield

Christian Edlich-Muth and Arren Bar-Even
Max Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam
edlich@mpimp-golm.mpg.de

For a new green revolution to sustainably feed the continually increasing world population, crop productivity will have to be significantly improved. Photorespiration represents a big challenge in this respect, because it dissipates energy and leads to the futile loss of CO_2, thereby limiting plant growth yield. The futureAgriculture consortium is proposing to replace native photorespiration with synthetic shunts that fulfill the same function but avoid the most important defect of photorespiration: the counterproductive release of CO_2.

Having identified the release of carbon as the main defect of photorespiration, we designed synthetic pathways that could replace the native pathway. We propose two alternatives, a carbon-neutral and a carbon-fixing shunt. To demonstrate the advantages of our pathway, we employ computational modeling of the synthetic shunts operating alongside the Calvin cycle. We use two approaches, firstly a hybrid stoichiometric-kinetic model and secondly a fully kinetic model. The former includes Rubisco kinetics, CO_2 diffusion and a simplified model of the light reactions. All other reactions are not explicitly modeled and are implied by stoichiometric relations between fluxes. The latter explicitly models every reaction of the shunt with reversible Michaelis-Menten kinetics.

The hybrid model has the advantage that it paints the full picture of photosynthesis with all processes involved. The performance of native photorespiration as compared to the synthetic shunts are evaluated in a variety of conditions that include the agriculturally most relevant ones, namely high versus low light exposure in combination with low availability of CO_2. The models clearly show that our synthetic shunts have a considerable advantage over native photorespiration in every imaginable scenario, with improvements in yield (fixed carbon) of up to 60%. The analysis also very strongly supports our initial hypothesis that CO_2 release is the main problem that needs to be tackled.

The hybrid model confirms the superiority of our synthetic shunts on a fundamental but relatively abstract level. However, it ignores the fact that a sufficient amount of activity of the synthetic enzymes needs to be available to support the fluxes. What is more, from the viewpoint of practical implementation, it offers no helpful insights. This is the strength of the kinetic model in which mission-critical enzyme properties and metabolite concentrations come under scrutiny. The most important outcome of these simulations is the prediction of which level of activity and specificity each enzyme in the pathway will be required to have. This information is invaluable for experimenters whose efforts in engineering novel enzymatic activities can now be directed towards specific goals. A second important outcome is the confirmation that the concentrations of metabolites in our pathways would remain within physiologically feasible limits.

In summary, our modeling of photorespiratory shunts has confirmed that these pathways have the potential to increase the photosynthetic efficiency of many if not most cultivated crops. Moreover, we are able to define the parameters within which an implementation of a synthetic pathway can be expected to be successful in the context of the chloroplast.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

MooViE: Multi-objective optimization Visualization Engine

Martin Beyß and Katharina Nöh
Institute of Bio- and Geosciences, IBG-1: Biotechnology, Forschungszentrum Jülich GmbH
ma.beyss@fz-juelich.de

In ^{13}C Metabolic Flux Analysis (^{13}*C-MFA*) microorganisms are fed with ^{13}C-labeled tracers and the labeling incorporation within the cells' metabolites is used to infer the intracellular reaction rates, called *metabolic fluxes*. A variety of tracers is available differing in both, the ability to resolve the fluxes and price. As prerequisite of ^{13}C-MFA an informative tracer mixture is to be selected, while, at the same time, budget constraints have to be obeyed. This gives rise to a multi-objective experimental design (*MO-ED*) problem that is further complicated by the availability of several different information measures [MWKdG99]. Different from single objective ED, MO-ED generates many compromise solutions for which none of the objectives can be improved without impairing the optimality of others. This notion of optimality is called *Pareto optimality*.

The MO-ED task in ^{13}C-MFA is high dimensional and the outcome inherently difficult to interpret. An effective visualization must facilitate understanding, i.e., allow for assessing the conflicts between the objectives, connecting the Pareto-optimal objective values to the related design settings, and supporting the selection process. Here, available approaches (e.g., scatterplots, parallel coordinates, dimension reduction) focus on displaying the objective values and suffer from various limitation [ALC+04].

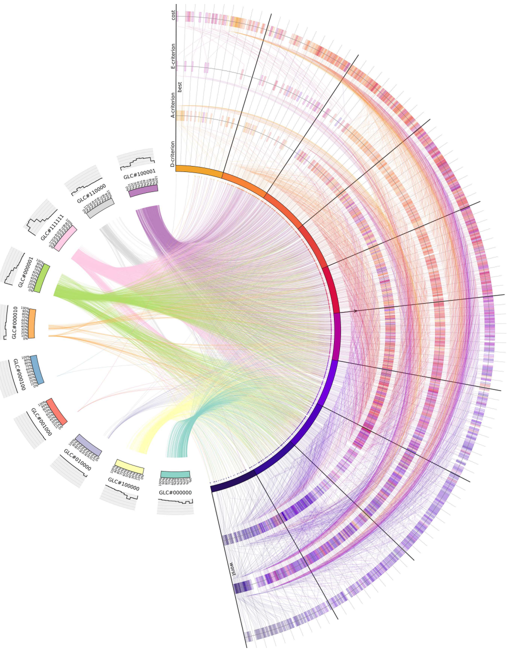


Figure 1: *MooViE* example for MO-ED: Over 1,000 Pareto sets with a 10-D tracer space (left) and a 4-D objective space (3 quality criteria and costs, right). The image is created with *circos* (www.circos.ca).

We introduce a novel approach for intuitive, information rich and appealing visualization of high-volume MO-ED results: *MooViE*. Color enhanced parallel coordinates in polar coordinates are extended by a chord diagram which displays the mapping from design variables to objective values via connecting arches. The use of established visualization elements enables easy comprehension. While the visualization is currently designed for providing static images, it has great potential for interactivity, to be used for in depth exploration. This will enable experimenters to chose the optimal tracer mixture for their specific ^{13}C-MFA use-case.

References

[ALC+04] G. Agrawal, K. Lewis, K. Chugh, C.H. Huang, S. Parashar, and C.L. Bloebaum. Intuitive visualization of Pareto frontier for multi-objective optimization in n-dimensional performance space. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, volume 30, 2004.

[MWKdG99] M. Möllney, W. Wiechert, D. Kownatzki, and A.A. de Graaf. Bidirectional reaction steps in metabolic networks: IV. Optimal design of isotopomer labeling experiments. *Biotech. and Bioeng.*, 66(2):86–103, 1999.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

MinHash Protein Similarity with Variable Window Size

Henning Timm, Sven Rahmann
Genome Informatics, Institute of Human Genetics,
University Hospital Essen, University of Duisburg-Essen, Germany
henning.timm@uni-due.de

The main goal of metagenomic sequencing is the identification of organisms and functions within a sample. This can be achieved by assigning the reads to representative protein sequences. However, finding the most similar protein sequence for a read in a large reference database poses several challenges. A read can be translated into up to six protein sequences, but given a DNA read length of about 100 bp they only have a length of approximately 30 aa. Additionally, protein sequences vary greatly in length, ranging from lengths in the tens of characters to ten-thousands. This can be problematic for similarity measures like the Jaccard index. Finally, reads might differ greatly from all reference sequences, if no sequence of the sampled organism is present in the database. Established tools like RAPSearch [YCT11, ZTY11] employ a seed-and-extend approach using short protein k-mers and a BLAST-like alignment algorithm.

We propose a flexible approach to indexing the reference using winnowed minimizers and min-hashing [SWA03, RHH+04] to identify alignment candidate sequences. By partitioning the reference sequences into windows of variable width, according to their minimizers, we solve the problem of the high variance in reference sequence length and the potential length difference between reads and references. We use s different hash functions for the partitioning to ensure that local minima of repetitive regions do not dominate the analysis. Hence, by choosing a different sketch size s, we can directly influence the trade-off between performance (using only a few hash functions) and sensitivity (choosing a large sketch size s).

For each hash function, we create a hash table mapping minimizers to window information compressed into 32 bit integer values. These include the reference sequence, the target chunk in that sequence, and a flag to exclude highly repetitive k-mers. While we sacrifice precision for each singular hash function by only specifying a chunk of the target sequence, this information is reconstructed using the results of all s mappings. We use a cache-aware hash table implementation that further benefits from the small size of the entries, as more potential hits fit into the same cache page, resulting in fewer cache misses during the hash table lookups.

References

[RHH+04] Michael Roberts, Wayne Hayes, Brian R Hunt, Stephen M Mount, and James A Yorke. Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20(18):3363–3369, 2004.

[SWA03] Saul Schleimer, Daniel S Wilkerson, and Alex Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85. ACM, 2003.

[YCT11] Yuzhen Ye, Jeong-Hyeon Choi, and Haixu Tang. RAPSearch: a fast protein similarity search tool for short reads. *BMC bioinformatics*, 12(1):159, 2011.

[ZTY11] Yongan Zhao, Haixu Tang, and Yuzhen Ye. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126, 2011.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Efficient large-scale whole genome alignments based on the SuperGenome

André Hennig and Kay Nieselt
Integrative Transcriptomics, University of Tübingen
andre.hennig@uni-tuebingen.de

Next-generation sequencing technologies have opened the possibility for large genomic projects that aim to identify variations between genomes. For comparative analysis approaches, whole genome alignments (WGAs) play an important role. The runtime for all aligners increases at least quadratically with the number of genomes, making alignments of thousands of genomes - even for small genomes such as those of bacteria - not feasible.

Here, we present an approach to compute WGAs that is many folds faster than the conventional methods. The general idea is to efficiently combine smaller sets of aligned genomes to the full WGA. For the combine-step, we first compute the consensus sequence of each subset, their respective consensus sequences are then aligned, transfered into the coordinate-based SuperGenome [HJBN12] data structure and the different coordinate systems are converted into a common one, which represents the WGA that contains all genomes.

The current implementation uses `progressiveMauve` [DMP10] for the genome alignment. To compare the results of our approach with the standard `progressiveMauve`-generated WGA we used the pairwise consistency over all pairs of genomes in the alignment. Our first results on *S. aureus* genomes show a runtime decrease by a factor of 100 with a consistency that is negligibly worse than the one of the original `progressiveMauve`-generated WGA.

References

[DMP10] A.E. Darling, B. Mau, and N.T. Perna. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one*, 5(6):e11147, 2010.

[HJBN12] A. Herbig, G. Jäger, F. Battke, and K. Nieselt. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*, 28(12):i7–i15, 2012.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

ASMAC: Approximate String Matching With Aho-Corasick

Chris Bielow* and Sandro Andreotti*
Bioinformatics Solution Center, Institut für Informatik, Freie Universität Berlin
mail@bsc.fu-berlin.de

The Aho-Corasick (AC) algorithm [AC75] is a competitive trie-based exact string matching solution, allowing to search multiple patterns in an input text simultaneously. We have extended AC to allow for matching an arbitrary number of ambiguous elements in the input text during the search phase which enables approximate string matching with Aho-Corasick (ASMAC).

Matching of ambiguous characters is achieved by an augmented trie traversal with a parallel exploration of all subtrees representing a possible resolution of the ambiguous character in the input text. Compared to index-based methods, AC is very competitive for certain applications due to fast linear-time trie construction, memory efficiency and low disk I/O requirements.

In the well-known peptide-protein mapping problem, protein databases (input text) commonly contain ambiguous amino acids which are resolved by peptide search engines when reporting peptide hits (dictionary) from measured mass spectral data. Common ambiguous amino acids are X (any amino acid), B (Aspartic acid or Asparagine), Z (Glutamic acid or Glutamine), I (Leucin or Isoleucin). We benchmark our implementation of ASMAC against index-based algorithms for approximate string matching in terms of speed and memory consumption for a variety of input data, where dictionary and input text size are varied according to realistic conditions, i.e. large protein databases for metagenomics or large dictionaries for high-complexity samples from a single organism.

Our algorithm is based on the AC implementation of SeqAn [DWRR08] and is available open-source in the PeptideIndexer tool as part of the OpenMS software suite [RSA$^+$16].

References

[AC75] Alfred V. Aho and Margaret J. Corasick. Efficient String Matching: An Aid to Bibliographic Search. *Commun. ACM*, 18(6):333–340, June 1975.

[DWRR08] Andreas Döring, David Weese, Tobias Rausch, and Knut Reinert. SeqAn An efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9(1):11, Jan 2008.

[RSA$^+$16] Hannes L Rost, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E Wolski, Oliver Schilling, Jyoti S Choudhary, Lars Malmstrom, Ruedi Aebersold, Knut Reinert, and Oliver Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Meth*, 13(9):741–748, September 2016.

*shared first authors

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Revitalizing alignment methods for efficient and accurate estimation of pairwise distances of divergent genomes

Annika Seidel and Stefan Kurtz
Universität Hamburg, MIN-Fakultät, ZBH - Center for Bioinformatics, Bundesstrasse 43, 20146 Hamburg, Germany, kurtz@zbh.uni-hamburg.de

The evolutionary distance between two, potentially long, homologous DNA sequences is used in diverse biological applications. Traditionally, the distances are inferred from pairwise sequence alignments. While this procedure is still standard in the mentioned application areas, approaches based on word statistics have become popular in recent years (cf. [Hau13]), due to the fact that current alignment-based methods do not scale for the ever growing size of the sets of genomes to be compared. The methods based on word statistics are often denoted as alignment-free methods, although some actually perform a gap-less alignment step, e.g. [HKP15]. The main problem with current alignment-free methods is the fact the distance estimates they deliver quickly become inaccurate for more divergent genomes.

We have developed an improved method for estimating pairwise genome distances based on techniques originally developed in [Mye14] and efficiently implemented in the `seed_extend` software developed as part of the GenomeTools (http://genometools.org) software package.

Our method uses a seed-extend approach in which seeds are extended if they occur in a diagonal band with a sufficient coverage of seeds and if they are not covered by the previously computed alignment. Both conditions, if appropriately parameterized very effectively reduce the number of the relatively costly seed-extensions. The extension itself is based on the greedy algorithm to compute the unit-edit distance of two strings with a heuristic trimming strategy to reduce the search space. For appropriate parameters, the extension step is expected to be linear in the length of the computed alignment. The identity values of a sufficient number of local alignments gives a substitution and indel rate and in turn an estimate of the overall genome distance.

Preliminary results: For simulated data as well as for pairs of real genome sequences with pairwise distances in the range from 1–30%, our method delivers distance values that almost perfectly correlate with the Average Nucleotide Identity (ANI), as computed by the dnadiff program of the MUMmer package [KPD+04]. Furthermore, estimations of the pairwise distances are accurate for ranges of up to 30% even in the presence of alignments with indels. This is an improvement over andi, which only counts mismatches and does not account for indels. For several benchmark sets of *E.coli*, *Brucella* and *Ebola* genomes we are able to accurately reconstruct established phylogenetic trees. Our method is faster by a factor of approx. $4-8$ and $1.15-3$ compared to dnadiff and andi [HKP15], respectively.

The poster will shortly describe the method and otherwise focus on the results of our method in comparison to previous methods for simulated and real data.

References

[Hau13] B. Haubold. Alignment-free phylogenetics and population genetics. *Briefings in bioinformatics*, 15(3):407–418, 2013.

[HKP15] B. Haubold, F. Klötzl, and P. Pfaffelhuber. `andi`: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31(8):1169–1175, 2015.

[KPD+04] S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg. Versatile and Open Software for Comparing Large Genomes. *Genome Biology*, 5(R12), 2004.

[Mye14] G. Myers. Efficient Local Alignment Discovery amongst Noisy Long Reads. In *Algorithms in Bioinformatics - 14th International Workshop, WABI 2014, Wroclaw, Poland, September 8-10, 2014. Proceedings*, pages 52–67, 2014.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

spalter: Separating genetic variants from sequencing errors based on technical features

Till Hartmann, Sven Rahmann
Genome Informatics, Institute of Human Genetics,
University Hospital Essen, University of Duisburg-Essen, Germany
till.hartmann@uk-essen.de

In DNA-sequencing data, mismatches with respect to a known reference sequence can either be explained by biological variance (e.g. SNVs) or technical errors (e.g. caused by flawed library preparation or sequencing bias). Being able to distinguish between these possibilities is crucial in order to avoid misclassification.

A common approach to achieve this lies within making assumptions about the distributions of biological variants and technical errors and ultimately performing hypothesis testing based on the obtained models [LCZ$^+$12, WWH$^+$11]. Since both distributions may vary across different species, sequencing techniques, sequencing machines or even due to library preparation, our goal was to remove the need to make explicit assumptions about distributions and error rates. The only assumption we do make is that there is a pattern to the occurrence of technical errors, i.e. technical errors occur *systematically*.

Many machine learning techniques are well suited for pattern recognition tasks, and in fact have been used to try to solve the SNP calling problem [OWDC13]. However, supervised machine learning algorithms need labeled data, which is rather limited in practice, while unsupervised algorithms (e.g. clustering) do not but are ill-suited for the task at hand. Our approach is based on pseudo-supervised machine learning and the concept of "learnability": For each locus, label bases with either of the two class labels "mismatch" or "match" w.r.t. the reference sequence and train a *simple* machine learning model. If the prediction of the respective model for that locus is accurate, a technical error is assumed. This has been implemented in our tool SPALTER.

We compare technical error calls made by SPALTER with SNV/SNP calls made by FREEBAYES [GM12] and show that likely technical errors correspond to low quality calls made by FREEBAYES and unlikely technical errors correspond to high quality calls made by FREEBAYES.

References

[GM12] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints*, July 2012.

[LCZ$^+$12] Bingshan Li, Wei Chen, Xiaowei Zhan, Fabio Busonero, Serena Sanna, Carlo Sidore, Francesco Cucca, Hyun M. Kang, and Gonçalo R. Abecasis. A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet*, 8(10):–1002944, 2012.

[OWDC13] Brendan D. O'Fallon, Whitney Wooderchak-Donahue, and David K. Crockett. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics*, 29(11):1361–1366, 2013.

[WWH$^+$11] Zhi Wei, Wei Wang, Pingzhao Hu, Gholson J. Lyon, and Hakon Hakonarson. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic acids research*, 39(19):–132, 2011.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

NGS Workflow for Minimal Residual Disease Follow-up

Razif Gabdoulline, Felicitas Thol, Piroska Clement, Johannes Schiller, Alessandro Liebich, Christian Kandziora, Madita Flintrop, Mira Pankratz, Martin Wichmann, Blerina Neziri, Arnold Ganser and Michael Heuser

Department of Hematology, Hemostasis, Oncology, and Stem Cell Transplantation; Integrated Research and Treatment Center Transplantation, IFB-Tx, Hannover Medical School, Hannover

gabdoulline.razif@mh-hannover.de

Approximately 30-40% of patients with acute myeloid leukemia (AML) develop a clinical relapse after allogeneic hematopoietic stem cell transplantation (alloHSCT). Minimal residual disease (MRD) testing may detect molecular signs of relapse prior to clinical relapse in patients without morphologic/microscopic evidence of disease. We set up a next-generation sequencing (NGS) workflow for the initial screening of mutations at diagnosis and for follow-up during the course of the treatment including alloHSCT; the workflow is already being applied to several hundreds of patient cases. The initial screening is based on the True Sight Myeloid Sequencing Panel targeting 54 candidate genes. We implemented appropriate state of the art NGS data analysis tools (alignment with bwa mem, base quality recalibration and indel realignment with GATK, variant calling with GATK and LoFreq, structural variant detection with Pindel) and developed filters to remove false positives. The filters are based on the analysis of (1) the PCR amplification process, (2) read positioning in the panel design, (3) overall sequencing quality of a sample, (4) co-occurrence of mutations in various cohorts of patients, and (5) the results of validation by Sanger sequencing. The analysis gives a realistic number of 2-3 mutations per patient on average; Sanger sequencing confirmed approximately 80% of mutations selected for validation. The workflow is automated, parallelized, optimized for speed and implemented on a linux virtual machine. Analysis time per sample is 6 CPU hours, permitting real-time processing of data on multiprocessor computers. Selected mutations are monitored by error-corrected amplicon sequencing with mutation detection sensitivity at the level of 0.01%. This sensitivity is only achievable, when typical sequencing errors of ca 0.15% are significantly reduced. For that we used a primer design, which labels (barcodes) short reads originating from the same DNA molecule. Resulting short reads are subjected to downstream computational correction removing variations accumulated during the whole sequencing process. Effectiveness of the method is tested by correcting for sequencer-only errors via combining forward and reverse reads, as well as by experimenting with polymerases from different suppliers. The tests show that the method reduces both sequencing and PCR amplification errors. Analysis takes less than 1 CPU hour, being applicable in clinical diagnostics. Applying this workflow we compared the cumulative incidence of relapse (CIR) in patients with positive and negative MRD. The CIR at 5 years was 67.7% in MRD positive patients and 19% in MRD negative patients ($P < 0.001$). Thus, we were able to separate patients with high and low risk of relapse after alloHSCT based on our NGS-MRD results. In conclusion, an NGS-MRD workflow can be established to achieve clinically meaningful results in a clinically relevant time frame.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Protein structure modeling with the assistance of workflow languages

Lukas Zimmermann, Luis de la Garza, and Oliver Kohlbacher
Applied Bioinformatics, University of Tübingen
lukas.zimmermann@informatik.uni-tuebingen.de

The specificity of mechanistic interactions between proteins is largely governed by their three-dimensional folding [1, 2]. Protein structure elucidation, which can be aided by *in silico* modeling, is thus vital for the understanding of molecular interactions [3].

Here we show that protocols for *in silico* protein structure modeling can be embodied as formal descriptions of the individual steps. As one way of doing so we chose the workflow representation of KNIME [4]. This software consists of a systematic way to denote and store workflows, and of a workflow engine for their execution. KNIME is already employed in cheminformatics applications and we seek to also benefit from KNIME in the field of structural bioinformatics [5]. As a proof-of-concept, we implemented two workflows, cross-link based comparative and *de novo* modeling, which have been described in detail by Kahraman *et al.* [6]. The challenge was to encode the required tools, such as ROSETTA, into a common format, such that they can subsequently be embedded into KNIME [7]. We used the Common Tool Description (CTD) format for this purpose, which allows the specification of program input/output and parameters in a platform-independent way [8]. Hence, CTD also facilitates the usage of the tools in any other workflow language than KNIME.

The advantages of such a workflow-centric approach are threefold. First, the workflow can be deposited as such in dedicated repositories like myExperiment [9]. Second, the reproducibility of the respective study is enhanced, for the individual steps are clearly documented and the published workflow ideally can be executed by a workflow engine with minimal effort from the user. Third, other researchers can easily carry on the research conducted in the original study, since the workflow can be adapted to new problem instances.

References

[1] Tony Pawson and Piers Nash. "Protein–protein interactions define specificity in signal transduction". In: *Genes & Development* 14.9 (May 2000), pp. 1027–1047.

[2] Qiangfeng Cliff Zhang et al. "Structure-based prediction of protein-protein interactions on a genome-wide scale". In: *Nature* 490.7421 (Oct. 2012), pp. 556–560.

[3] Patrick Aloy and Robert B Russell. "Structural systems biology: modelling protein interactions". In: *Nat Rev Mol Cell Biol* 7.3 (Mar. 2006), pp. 188–197.

[4] Michael R Berthold et al. "KNIME - the Konstanz Information Miner: Version 2.0 and Beyond". In: *SIGKDD Explor. Newsl.* 11.1 (Nov. 2009), pp. 26–31.

[5] Stephan Beisken et al. "KNIME-CDK: Workflow-driven cheminformatics". In: *BMC Bioinformatics* 14.1 (Aug. 2013), p. 257.

[6] Abdullah Kahraman et al. "Cross-Link Guided Molecular Modeling with ROSETTA". In: *PLOS ONE* 8.9 (Sept. 2013), e73411.

[7] Kristian W Kaufmann et al. "Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You". In: *Biochemistry* 49.14 (Apr. 2010), pp. 2987–2998.

[8] Marc Röttig. *Combining Sequence and structural Information into Predictors of enzymatic Activity.* München: Hut, 2013.

[9] C a. Goble et al. "myExperiment: a repository and social network for the sharing of bioinformatics workflows". In: *Nucleic Acids Res* 38 (2010).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

An algorithm to build a multi-genome reference

Leily Rabbani, Jonas Mueller, Detlef Weigel
Max Planck Institute for Developmental Biology
leily.rabbani@tuebingen.mpg.de

As a result of next generation sequencing technologies, during the last decade many studies have aimed to analyze thousands, and often many more genomes. Clearly, comparing these against only a single reference genome sequence is no longer adequate. To overcome the limits imposed by using a single reference genome, we introduce a method to represent several genomes in a simplified data structure, which allows simultaneous comparison against multiple genomes. To this end, we developed an algorithm that creates a graph as a multi-genome reference. To reduce the complexity of this representation, highly similar orthologous and paralogous regions are collapsed and only a representative piece of sequence is picked to present the similar regions. The algorithm aids in removing the bias against a single-genome reference and simplifies downstream analysis.

To evaluate the efficiency of our model, we developed a genome compression algorithm which is able to compress the entire input sequences. The superior compression rate confirms that our model fits sequence properties better than existing DNA compression programs. The compression algorithm can be also be used for global genome comparison.

In future, we anticipate that even with very cheap long reads, not every genome will be assembled de novo, once a sufficient number of platinum standard genomes is available. To identify differences from available assemblies, long reads should be mapped against a genome graph. We therefore also designed a mapping algorithm to search for the path on the graph that fits a long read the best.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

SubtiWiki, an integrated and robust database for model organism Bacillus subtilis

Bingyao Zhu and Jörg Stülke
Department of General Microbiology, University of Göttingen
Grisebachstr. 8, 37077 Göttingen, Germany
bzhu@gwdg.de

*Subti*Wiki [MZMS16] is a popular database for model organism *Bacillus subtilis substr. 168* with over 10,000 hits per day. It provides public access to all users without subscription or registration. It collects manually curated annotations about genes and new RNA features. In *Subti*Wiki, annotations of each gene/RNA feature are presented in a single webpage. In addition, we have created multiple browsers to visualize associations among genes. In our expression browser, transcriptomic and proteomic data are presented as interactive charts. The new genome browser provides researchers a quick look of the genomic context of a gene. Biochemical pathways in *Bacillus subtilis* are shown as maps in our pathway browser. In our interaction and regulation browser, biological networks namely protein-protein interaction and gene regulation are visualized using a gravity model. The nodes in the network are modelled as mass points with weight and edges as springs. Nodes repel each other as gravity constant is set to be negative while edges hold the nodes together. The layout of the network is determined by the Barnes-Hut simulation [BH86]. With our iOS and Android App (available in App Store and Google Play Store), mobile access to our data is guaranteed.

All data stored in *Subti*Wiki can be exported in parseable format for further analysis. Both the protein-protein interaction network and the gene regulation network can be exported as adjacency list in csv format.

We also provide an offline tool named *NetVis*, which applies the same model as in the regulation browser. Users can display networks downloaded from *Subti*Wiki or visualize their own networks. The appearance of the network can be individually adjusted. The displayed network can also be modified by adding/deleting/editing nodes and edges. The result of simulation can be saved or exported as image.

References

[BH86] J. E. Barnes and P. Hut. A hierarchical O(n-log-n) force calculation algorithm. *Nature*, 324:446, 1986.

[MZMS16] Raphael H. Michna, Bingyao Zhu, Ulrike Mäder, and Jörg Stülke. SubtiWiki 2.0, an integrated database for the model organism Bacillus subtilis. *Nucleic Acids Research*, 44(D1):D654–D662, 2016.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

seq-seq-pan: Rapid construction of a computational pan-genome data structure with iterative updates

Christine Jandrasits[1], Piotr W. Dabrowski[1], Stephan Fuchs[2] and Bernhard Y. Renard[1]
[1] *Bioinformatics Unit (MF1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin*
[2] *Department of Infectious Diseases, Robert Koch Institute, Wernigerode Branch, Wernigerode*
JandrasitsC@rki.de, RenardB@rki.de

The increasing application of next generation sequencing technologies has led to the availability of thousands of reference genomes, often providing multiple genomes for the same or closely related species. The current approach to represent a species or a population with a single reference sequence and a set of variations cannot represent their full diversity and introduces bias towards the chosen reference. There is a need for the representation of multiple sequences in a composite way that is compatible with existing data sources for annotation and suitable for established sequence analysis methods. At the same time, this representation needs to be easily accessible and extendable to account for the constant change of available genomes.

We introduce seq-seq-pan, a framework that provides methods for adding or removing new genomes from a set of aligned genomes and uses these to construct a whole genome alignment. Throughout the sequential workflow the alignment is optimized for generating a representative linear presentation of the aligned set of genomes, that enables its usage for annotation and in downstream analyses. We use progressiveMauve [DMP10], a whole genome aligner, that focuses on the detection of large sequence rearrangements, for pairwise alignments in each step and snakemake [KR12] for workflow management. We compared seq-seq-pan with three whole genome aligners that offer alignment of non-collinear sequences using simulated and real data.

By providing dynamic updates and optimized processing, our approach enables the usage of whole genome alignment in the field of pan-genomics. In addition, the sequential workflow can be used as a fast alternative to existing whole genome aligners.

seq-seq-pan is freely available at https://gitlab.com/groups/rki_bioinformatics

References

[DMP10] Aaron E. Darling, Bob Mau, and Nicole T. Perna. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PloS One*, 5(6), 2010.

[KR12] Johannes Köster and Sven Rahmann. Snakemake - a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

de.NBI Cloud-Site Tübingen

Maximilian Hanussek[1,2], Felix Bartusch[1,2], Oliver, Kohlbacher[2], Thomas Walter[1] and Jens Krüger[1]
[1] *High Performance and Cloud Computing Group, IT Center, University of Tübingen*
[2] *Applied Bioinformatics Group, WSI/ZBIT, University of Tübingen*
jens.krueger@uni-tuebingen.de

A cloud solution for bioinformaticians is being provided through the university compute center in Tübingen, including computing infrastructure, various workflow solutions to construct pipelines and tools for all bioinformatic areas. The cloud-site environment of Tübingen involves the following services:

- Infrastructure as a Service (IaaS): The cloud infrastructure can be accessed via various virtualization technologies such as virtual machines, Docker containers or Singularity containers. The user will be able to use computing resources via the UNICORE middleware to run resource consuming data analysis and simulation algorithms. Furthermore, users can set up web services in their virtualization environment and use the available resources for their applications.
- Platform as a Service (PaaS): Tools provided will be UNICORE and Galaxy workflow systems and frameworks such as BALL. The user will be able to use the existing frameworks and workflows and further customize it on its own.
- Software as a Service (SaaS): Different kind of software will be available to use it right out of the box and apply it to scientific data without any further development[NBK+12]. The preinstalled software will be provided as virtual machine images, Docker containers, Singularity containers, UNICORE or Galaxy workflows [AHG+16]. The cloud site of Tübingen is focused, among other things, on the reproducibility of research data and their virtual research environments as Tübingen is a project partner of the CiTAR project.

One major aim of the de.NBI cloud site in Tübingen is to provide software, covering different scientific fields of research such as mass spectrometry analysis, NGS analysis pipelines but also molecular docking via Ball integrated into Galaxy workflows (ballaxy) [HSF+15].

The de.NBI cloud infrastructure in Tübingen comprises more than 1650 compute cores, 15 TByte RAM, 17 TByte SSD storage and 180 TByte storage capacity. The process to get access to cloud is relatively simple. For the time being, the manager of a project, which wants to use cloud resources, sends an application to the corresponding contact person (jens.krueger@uni-tuebingen.de), including the information of what is the project about, how many resources are needed (CPUs, RAM, Storage, IPs), how long does the project run and which people need access to the cloud. Other specifications can individually be discussed. If the application is accepted you will get access to your own cloud project and be able to launch and administrate virtual machines (VMs) via the Openstack Dashboard or the Openstack API. The started VMs can securely be accessed by SSH-Keys.

References

[AHG+16] Junaid Arshad, Alexander Hoffmann, Sandra Gesing, Richard Grunzke, Jens Krüger, Tamas Kiss, Sonja Herres-Pawlis, and Gabor Terstyanszky. Multi-level meta-workflows: new concept for regularly occurring tasks in quantum chemistry. *Journal of Cheminformatics*, 8(1):58, Oct 2016.

[HSF+15] Anna Katharina Hildebrandt, Daniel Stöckel, Nina M. Fischer, Luis de la Garza, Jens Krüger, Stefan Nickels, Marc Röttig, Charlotta Schärfe, Marcel Schumann, Philipp Thiel, Hans-Peter Lenhof, Oliver Kohlbacher, and Andreas Hildebrandt. ballaxy: web services for structural bioinformatics. *Bioinformatics*, 31(1):121–122, 2015.

[NBK+12] Oliver Niehörster, André Brinkmann, Axel Keller, Christoph Kleineweber, Jens Krüger, and Jens Simon. Cost-Aware and SLO-Fulfilling Software as a Service. *Journal of Grid Computing*, 10(3):553–577, Sep 2012.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3268v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017