Poster abstracts of GCB 2015

Martin Eisenacher[1], Jörg Rahnenführer[2], Axel Mosig[3], and Sven Rahmann[4]

[1]**Medical Proteome Center, Ruhr University Bochum, Germany**
[2]**Department of Statistics, TU Dortmund University, Dortmund, Germany**
[3]**Bioinformatics, Department of Biophysics, Ruhr University Bochum, Germany**
[4]**Genome Informatics, Human Genetics, University Hospital Essen, University of Duisburg-Essen, Essen, Germany**

EDITORIAL

The poster session is and always has been an important part of the German Conference on Bioinformatics, as it gives an excellent overview of current bioinformatics research topics in Germany and other countries.
At GCB 2015, there will be 57 posters on display, whose abstracts (together with the poster number) are collected in the present document. Poster prizes will be presented to the presenting author(s) of the best posters, as selected by the program committee.

Keywords: GCB 2015, posters, abstracts

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

List of Posters

Comparison of variable importance measures from random forest algorithms on strongly unbalanced data

Ursula Neumann, Mona Riemenschneider and Dominik Heider
Department of Bioinformatics, Straubing Center of Science
u.neumann@wz-straubing.de

Background:
In bioinformatics and related research fields, variable importance measures (VIMs) have drawn increased attention in the last years. The purpose of VIMs is to identify a minimal subset of features (e.g. genetic markers) which are relevant for an intended prediction model.

There are different implementations of the random forest algorithm in R which offer diverse variable importance measures. Breiman's random forest [Bre01], available in the R package *randomForest* implemented by Liaw and Wiener [LW02], contains two VIMs, namely the error-rate-based and the Gini-index-based VIM. The cforest method from the *party* package, implemented by Hothorn et al.[HHZ06], provides an alternative error-rate-based and an AUC-based VIM.

Results:
We examined the four different VIMs on a dataset consisting of 437 patients, who had a myocardial infarction. We aimed to analyze a potential predictive value out of liver serum markers for the severity of stenosis in acute myocardial infarction. The results from the Gini-index-based VIM show clearly a preference in importance of those features which are not dichotomous. Whereas the other two VIMs reveal different important features for prediction of stenosis, including two dichotomous variables, namely *family predisposition* (mean of classes $\mu = 0.3$) and *dyslipidemia* (mean of classes $\mu = 0.2$).

Conclusion:
Dichotomous predictors associated with the target variable are discriminated in the importance analysis. These results support the findings of previous studies. Kononenko [Kon95] demonstrated that the values of the Gini-index measure increase linearly with the number of classes. Thus, the features family predisposition and dyslipidemia are discriminated compared to other features, because of their lower number of categories. There is a second reason for the discrimination of these two features by the Gini-index-based VIM: the unbalanced sizes of classes. Janitza et al. [JSB13] showed that the AUC-based permutation VIM has the best performance for models with varying numbers of samples in each class.

References

[Bre01] L Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[HHZ06] T Hothorn, K Hornik, and A Zeileis. party: A Laboratory for Recursive Part(y)itioning, 2006.

[JSB13] S Janitza, C Strobl, and AL Boulesteix. An AUC-based Permutation Variable Importance Measure for Random Forests. *BMC Bioinformatics*, 14:119, 2013.

[Kon95] I Kononenko. On Biases in Estimating Multi-Valued Attributes. *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada*, pages 1034–1040, 1995.

[LW02] A Liaw and M Wiener. Classification and Regression by randomForest. *R News*, 2:18–22, 2002.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Structure prediction of GPCRs using piecewise homologs and application to the human CCR5 chemokine receptor: validation through agonist and antagonist docking

Karthik Arumugam, Andy Chevigne, Carole Seguin-Devaux and Jean-Claude Schmit
Luxembourg Institute of Health, Department of Infection and Immunity, 29, rue Henri Koch, L-4354 Esch-sur-Alzette Luxembourg
karthik.arumugam@lih.lu

Our research describes the construction and validation of a three-dimensional model of the human CC chemokine receptor 5 (CCR5) receptor a GPCRs protein using multiple homology modeling. A new methodology is presented where we built each secondary structural model of the protein separately from distantly related homologs of known structure. The reliability of our approach for G-protein coupled receptors was assessed through the building of the human C-X-C chemokine receptor type 4 (CXCR4) receptor of known crystal structure. The models are refined using molecular dynamics simulations and energy minimizations using CHARMM, a classical force field for proteins. Finally, docking models of both the natural agonists and the antagonists of the receptors CCR5 and CXCR4 are proposed. This study explores the possible binding process of ligands to the receptor cavity of chemokine receptors at molecular and atomic levels. We proposed few crucial residues in receptors binding to agonist/antagonist for further validation through experimental analysis. In particular, our study provides better understanding of the blockage mechanism of the chemokine receptors CCR5 and CXCR4, and may help the identification of new lead compounds for drug development in HIV infection and other inflammatory diseases.

Poster 2

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

The Roles of Post-translational Modifications in the Context of Protein Interaction Networks

Guangyou Duan and Dirk Walther
Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm
walther@mpimp-golm.mpg.de

Among other effects, post-translational modifications (PTMs) have been shown to exert their function via the modulation of protein-protein interactions. For twelve different main PTM-types and associated subtypes and across 9 diverse species, we investigated whether particular PTM-types are associated with proteins with specific and possibly strategic placements in the network of all protein interactions by determining informative network-theoretic properties. Proteins undergoing a PTM were observed to engage in more interactions and positioned in more central locations than non-PTM proteins. Among the twelve considered PTM-types, phosphorylated proteins were identified most consistently as being situated in central network locations and with the broadest interaction spectrum to proteins carrying other PTM-types, while glycosylated proteins are preferentially located at the network periphery. For the human interactome, proteins undergoing sumoylation or proteolytic cleavage were found with the most characteristic network properties. PTM-type-specific protein interaction network (PIN) properties can be rationalized with regard to the function of the respective PTM-carrying proteins. For example, glycosylation sites were found enriched in proteins with plasma membrane localizations and transporter or receptor activity, which generally have fewer interacting partners. The involvement in disease processes of human proteins undergoing PTMs was also found associated with characteristic PIN properties. By integrating global protein interaction networks and specific PTMs, our study offers a novel approach to unraveling the role of PTMs in cellular processes.

Structural determinants of metabolite-protein binding events

Paula Korkuc and Dirk Walther
Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm
walther@mpimp-golm.mpg.de

To better understand and ultimately predict both the metabolic activities as well as the signaling functions of metabolites, a detailed understanding of the physical interactions of metabolites with proteins is highly desirable. Focusing in particular on protein binding specificity vs. promiscuity, we performed a comprehensive analysis of the structural determinants of metabolite-protein binding events as reported in the Protein Data Bank (PDB). We compared the molecular and structural characteristics obtained for metabolites to those of the well-studied interactions of drug compounds with proteins. Promiscuously binding metabolites and drugs are characterized by low molecular weight and high structural flexibility. Unlike reported for drug compounds, low rather than high hydrophobicity appears associated, albeit weakly, with promiscuous binding for the metabolite set investigated in this study. Across several physicochemical properties, drug compounds exhibit characteristic binding propensities that are distinguishable from those associated with metabolites. Compound properties capturing structural flexibility and hydrogen-bond formation descriptors proved particularly informative in PLS-based prediction models of binding promiscuity. With regard to diversity of enzymatic activities of the respective metabolite target enzymes, the metabolites benzylsuccinate, hypoxanthine, trimethylamine N-oxide, oleoylglycerol, and resorcinol showed very narrow process involvement, while glycine, imidazole, tryptophan, succinate, and glutathione were identified to possess broad enzymatic reaction scopes. Promiscuous metabolites were found to mainly serve as general energy currency compounds, but were identified to also be involved in signaling processes and to appear in diverse organismal systems (digestive and nervous system) suggesting specific molecular and physiological roles of promiscuous metabolites.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Solving the Differential Peak Calling Problem

Manuel Allhoff[1,2,3,*] Kristin Seré[3], Martin Zenke[3] and Ivan G. Costa[1,2,3]
[1] *Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Germany,*
[2] *Interdisciplinary Centre for Clinical Research (IZKF), RWTH University Medical School, Aachen, Germany,*
[3] *Helmholtz Institute for Biomedical Engineering, RWTH University Medical School, Aachen, Germany*
allhoff@aices.rwth-aachen.de

Identification of changes in DNA-protein interactions from Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) data is an important step to unravel regulatory biological processes. The differential peak calling problem is about finding genomic regions with changes in ChIP-seq signals describing the interaction of a protein with DNA between two cellular conditions. Several approaches, so-called two-stage differential peak callers, identify such genomic regions by using a combination of single peak callers with statistical tests for detecting differential digital expression. These two-stage differential peak callers fail to detect subtle changes of protein-DNA interactions. One-stage differential peak callers are based on signal segmentation strategies.

We propose a Hidden Markov Model based one-state differential peak callers: ODIN [ASC^{+}14] (One-stage DIffereNtial peak caller) finds differentials peaks in pairs of ChIP-seq data, whereas THOR is able to take replicates of ChIP-seq experiments into account. Both tools perform genomic signal processing, peak calling and p-value calculation in an integrated framework. We also propose an evaluation methodology to compare ODIN and THOR with competing methods. The evaluation is based on the association of differential peaks with expression changes in the same cellular conditions as well as simulated data. Our empirical study based on several ChIP-seq experiments from transcription factors, histone modifications and simulated data shows that our approaches perform better in most scenarios compared to the considered competing methods.

References

[ASC^{+}14] Manuel Allhoff, Kristin Seré, Heike Chauvistré, Qiong Lin, Martin Zenke, and Ivan G. Costa. Detecting differential peaks in ChIP-seq signals with ODIN. *Bioinformatics*, 30(24):3467–3475, 2014.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Prediction of interaction-dependent enzyme activity in the trisporic acid pheromone system by modelling of *in silico* mutants

Sabrina Ellenberger[1], Stefan Schuster[2], and Johannes Wöstemeyer[1]
[1] *Chair of General Microbiology and Microbial Genetics, Friedrich Schiller University Jena, 07743 Jena, Neugasse 24, Germany*
[2] *Department of Bioinformatics, Friedrich Schiller University Jena*
07743 Jena, Ernst-Abbe-Platz 2, Germany
Sabrina.Ellenberger@uni-jena.de

4-Dihydromethyltrisporate dehydrogenase (TSP1) is one of the enzymes of the trisporic acid biosynthesis pathway in zygomycetous fungi. This NADP-dependent aldo-keto reductase is acting on trisporoids, special pheromones, which are important for recognition of complementary mating types and sexual spore formation. In parasitic zygomycetes, these substances have an additional function. They are also responsible for host-parasite interactions and the formation of infection structures. We were interested in protein structures of TSP1 from *Parasitella parasitica*, the genome of which was recently sequenced by us, and in the mechanisms involved in the switch from sexual to parasitic communication which seems to be regulated at the protein level. Against expectation, *P. parasitica* contains six TSP1-like enzymes. Models of tertiary structures of the isoforms were created and compared with predicted structures of *in silico* mutants. We performed protein-protein docking with the resulting protein structures to simulate dimerization of wild type and mutants.

All TSP1 isoforms show the typical structure of aldo-keto reductases. Substrate specificity is determined by variations in the loops at the C-terminal side of the barrel. The comparative modelling approach reveals highly conserved binding pockets for the cosubstrate NADP in the different TSP1-like proteins, while the binding sites for the trisporoid substrate exhibit a higher degree of variation. TSP1 enzymes of different fungi are assumed to diversify due to preferring different trisporoid derivatives as sexual pheromones. Therefore, it appears reasonable that a fungus like *P. parasitica* which communicates with completely different partners during sexual or parasitic interactions, will keep a set of different enzymes in its trisporic acid communication system to respond differentially to its possible partners. TSP1 PARPA_07791 forms inactive homodimers and mediates the sexual pathway probably as in other zygomycetous fungi like *Mucor mucedo*. The second TSP1, PARPA_04105, forms active homodimers and could be responsible for the parasitic pathway of communication.

Three structure elements could be identified, that are needed for dimerization. These are i. the loop between the $\beta4$ strand and the $\alpha4$ helix (Fig1: green parts), ii. the $\alpha5$ helix (Fig1: blue parts), and iii. the loop region at the C-termini of the proteins (Fig1: yellow parts). Modification of three amino acids in the first region is sufficient to turn the inactive isoform into an active one (Fig1B). Inactivation of the enzyme is achieved by closure of the active site (Fig1: red parts) by the internal loop region of the binding partner and could not be induced by our mutants.

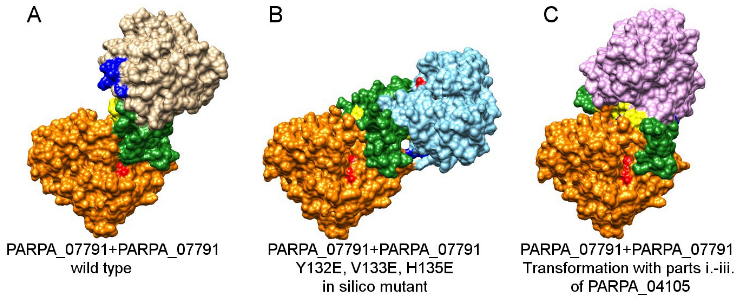

Figure 1: Activation of TSP1 PARPA_07791 homodimer visualized by modelled *in silico* mutants.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Pathway Analysis of a Herbicide Resistant Grass

Norma J. Wendel and Antje Krause
Fachhochschule Bingen
n.j.wendel@fh-bingen.de

Herbicide resistance is grouped in the well studied target site resistance (TSR) and the non-target site resistance (NTSR). The quantitative NTSR belongs to the reactions of abiotic stresses. It is a common issue in agriculture, but the biological background is largely unknown [Délye et al., 2013, Délye, 2013]. The non-model grass weed *Alopecurus myosuroides* (ALOMY) is resistant to many herbicides and shows TSR and NTSR. It is widespread in Western Europe and damages especially winter crops. ALOMY has almost non public sequence data available.

In our project the NTSR of ALOMY was studied on a systems biology and visualisation level. The goal was to explore metabolic pathways for the identification of genes or pathways involved in NTSR. To reach this goal reference species were needed with public sequence data (Ensembl) and pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG).

Time-series (before and after herbicide treatment) transciptome sequence data [Höfer et al., 2014] of the resistant ALOMY were mapped to nine reference species. The results were further pre-processed with SAMtools [Li et al., 2009] and Perl [Larry Wall, 2014] scripts. Limma [Smyth, 2005] generated Venn diagrams. They show the amount of genes mapped to the reference species for all time points and their intersections. The software package Pathview [Luo and Brouwer, 2013] performed the mapping to metabolic pathways of three reference species *Arabidopsis thaliana*, *Oryza sativa*, and *Sorghum bicolor*. This project provides the final results of this work including nine Venn diagrams with the underlying transcriptome data and 339 metabolic (KEGG) pathways.

Now further analysis have to be done. First, the identified transcripts and corresponding genes have to be compared and functionally observed. Pseudogenes and non-protein coding genes should be taken into account. Second, the pathways need to be studied in detail.

References

[Délye, 2013] Délye, C. (2013). Unravelling the genetic bases of non-target-site-based resistance (NTSR) to herbicides: a major challenge for weed science in the forthcoming decade. *Pest Management Science*, 69(2):176–187.

[Délye et al., 2013] Délye, C., Jasieniuk, M., and Le Corre, V. (2013). Deciphering the evolution of herbicide resistance in weeds. *Trends in Genetics*, 29(11):649–658.

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

[Larry Wall, 2014] Larry Wall (2014). About Perl. [Online; accessed: 18 December 2014; last update: 18 December 2014].

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

[Luo and Brouwer, 2013] Luo, W. and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831.

[Smyth, 2005] Smyth, G. K. (2005). Limma: linear models for microarray data. In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York.

Poster 7

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Identifying Sequences by k-Mer Analysis of Amino Acids in a Custom-Built Database

Silvio Weging and Andreas Gogol-Döring
German Center of Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany
Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany
silvio.weging@idiv.de

The usual way to process high-throughput sequencing data involves the alignment of the resulting reads against reference sequences. Unfortunately, fast read mapping tools like Bowtie [LTP+09] or Segemehl [HOK+09] are not designed to map reads with a high error tolerance or against very comprehensive sequence data bases like the NCBI nr/nt database [PTM07] as it would be desirable when analyzing metagenomic sequencing data. Alignment tools capable to perform this task like Blast [AGM+90], on the other hand, are relatively slow, which turns the mapping procedure into a computationally demanding task. One strategy for decreasing the amount of sequences to be mapped is to assemble the reads into longer contigs. Sequence assembly on the other hand is also a very complex and error prone task potentially generating chimeric sequences or bluring subtle sequence variations [FB09].

To overcome this problem we applied a method that bases on indexing k-mers [WS14], i.e., substrings of length k, and we developed a highly efficient algorithm in C++ for comparing k-mer indices generated on sequencing reads with previously compiled databases containing the k-mer frequencies within branches of the taxonomic tree. The user may pre-compile this database for faster comparisons while still leaving the choice of which entries shall be compared to the data. If the user does not know the possible taxonomic affiliations of the sample, the level of abstraction in the taxonomic tree can be chosen arbitrarily and refined iteratively. Our method indexes amino acids instead of DNA because they are better conserved and are therefore more robust against mutation and variation such that our software can assign sequences of diverse organisms much better than software running wholly on DNA [EOD+12]. This also results in smaller index sizes and therefore improves the performance of our tool. Because of the size of the given data, we decided to utilize the hard disk as an extension to the memory using the C++ STXXL library [RD05]. This leads to the possibility of running our tool on a desktop computer.

References

[AGM+90] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[EOD+12] Robert A Edwards, Robert Olson, Terry Disz, Gordon D Pusch, Veronika Vonstein, Rick Stevens, and Ross Overbeek. Real Time Metagenomics: Using k-mers to annotate metagenomes. *Bioinformatics*, 28(24):3316–3317, 2012.

[FB09] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6:6–12, 2009.

[HOK+09] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, 2009.

[LTP+09] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biol*, 10(3):R25, 2009.

[PTM07] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1):D61–D65, 2007.

[RD05] Peter Sanders Roman Dementiev, Lutz Kettner. Stxxl: Standard Template Library for XXL Data Sets. Technical Report 18, Fakultät für Informatik, Universität Karlsruhe, 2005.

[WS14] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

SHIVA - A web application for drug resistance testing in HIV

Mona Riemenschneider, Thomas Hummel, Ursula Neumann, Dominik Heider
Department of Bioinformatics, Straubing Center of Science, Straubing, Germany
m.riemenschneider@wz-straubing.de

Drug resistance testing is mandatory in antiretroviral therapy (ART) in human immunodeficiency virus (HIV) infected patients for successful treatment. The emergence of resistances against antiretroviral agents remains the major obstacle in inhibition of viral replication and thus to control infection. Due to the high mutation rate the virus is able to adapt rapidly under drug pressure leading to the evolution of resistant variants and finally to therapy failure.

Here we provide a web service for drug resistance prediction of commonly used drugs in ART, i.e. protease inhibitors (PIs) and reverse transcriptase inhibitors (NRTIs and NNRTIs), but also of novel drugs, such as maturation inhibitors. Furthermore, co-receptor tropism (CCR5 or CXCR4) can be predicted as well, which is essential for treatment with entry inhibitors, such as Maraviroc. SHIVA is able to integrate new computational models very easily via an XML configuration file. Currently, it provides 17 models [HSCH13, HDWH14, DRH^{+}11] for several drug classes. SHIVA can be used with single RNA or amino acid sequences, but also with huge amounts of data from next-generation sequencing and allows prediction of a user specified selection of drugs simultaneously, providing results as clinical reports via email to the user.

References

[DRH^{+}11] J Nikolaj Dybowski, Mona Riemenschneider, Sascha Hauke, Martin Pyka, Daniel Hoffmann, and Dominik Heider. Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers. *BioData Mining*, 4:26, 2011.

[HDWH14] Dominik Heider, Jan Nikolaj Dybowski, Christoph Wilms, and Daniel Hoffmann. A simple structure-based model for the prediction of HIV-1 co-receptor tropism. *BioData Mining*, 7:14, 2014.

[HSCH13] Dominik Heider, Robin Senge, Weiwei Cheng, and Eyke Hüllermeier. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*, 29(16):1946–1952, 2013.

Poster 9

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

GNATY: A tools library for faster variant calling and coverage analysis

Beat Wolf[1,2], Pierre Kuonen[1], Thomas Dandekar[2]
[1] *University of Applied Sciences Western Switzerland, Fribourg*
[2] *University of Würzburg, Germany*
beat.wolf@hefr.ch

Following the speed increases in next generation sequencing over recent years, the proportion of time spent in sequence analysis compared to sequencing has increasingly shifted towards sequence analysis. While certain analysis steps such as sequence alignment were able to benefit from various speed increases, others, equally important steps like variant calling or coverage analysis, did not receive the same improvements. Analysing NGS data remains a complicated and time consuming process, requiring a substantial amount of computing power. Most current approaches to address the increasing data quantity rely on the usage of more powerful hardware or offload calculations to the cloud. In this poster we show that by using modern software development techniques such as stream processing, those additional analysis steps can be sped up without changing the analysis results. Developing more efficient implementations of existing algorithms makes it possible to process larger datasets on existing infrastructure, without changing the analysis results. This not only reduces the overall cost of data analysis, but also gives researches more flexibility when exploring different settings for the data analysis. We present the application GNATY, a stand-alone version of NGS data analysis tools used in Gensearch-NGS [WKDD15] developed by Phenosystems SA. The goal during the development of the GNATY tools was not to create new methods with different results to existing approaches, but explore the possibilities of improving the efficiency of existing approaches. A modular architecture has been developed to create efficient sequence alignment analysis tools, using stream processing techniques which allow for multithreading and reusable data analysis blocks. The modular architecture uses a stream processing based workflow, efficiently splitting data access and data processing analysis steps, resulting in a more efficient use of the available computing resources. The architecture has been verified by implementing a variant caller based on the Varscan 2 [KZL$^+$12] variant calling model, achieving a speedup of nearly 18 times. The results of the variant calling in GNATY are identical to Varscan 2, avoiding the issue of adding yet another variant calling model to the existing ones. To further demonstrate the flexibility and efficiency of the approach, the algorithm is also applied to coverage analysis. Compared to BEDtools 2 [QH10], GNATY was twice as fast to perform coverage analysis, while producing the exact same results.

Through the example of 2 existing next generation sequencing data analysis algorithms which are reimplemented with an efficient stream based modular architecture, we show the performance potential in existing data analysis tools. We hope that our work will lead to more efficient algorithms in bioinformatics in general, lessening the hardware requirements to cope with the ever increasing amounts of data to be analysed. The developed GNATY software is freely available for non-commercial usage at http://gnaty.phenosystems.com/.

References

[KZL$^+$12] Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012.

[QH10] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[WKDD15] B. Wolf, P. Kuonen, T. Dandekar, and Atlan D. DNAseq Workflow in a Diagnostic Context and an Example of a User Friendly Implementation. *BioMed Research International*, 2015:11, 2015.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Robust Signature Selection from High-Throughput Genomic Data

Sangkyun Lee[1,*], Jörg Rahnenführer[2], Michel Lang[2], Katleen De Preter[3], Pieter Mestdagh[3], Jan Koster[4], Rogier Versteeg[4], Raymond L. Stallings[5], Luigi Varesio[6], Shahab Asgharzadeh[7], Johannes H. Schulte[8−12], Kathrin Fielitz[8], Melanie Schwermer[8], Katharina Morik[1], Alexander Schramm[8]

[1] *Department of Computer Science, TU Dortmund University, Germany* [2] *Department of Statistics, TU Dortmund University, Germany* [3] *Center for Medical Genetics, Ghent University Hospital, Belgium* [4] *Department of Oncogenomics, Academic Medical Center, Amsterdam, the Netherlands* [5] *Cancer Genetics, Royal College of Surgeons, Dublin, Ireland* [6] *Laboratory of Molecular Biology, Giannina Gaslini Institute, Genova, Italy* [7] *Hematology/Oncology, Children's Hospital Los Angeles, Los Angeles, USA* [8] *Department of Pediatric Oncology and Hematology, University Children's Hospital Essen, Germany* [9] *Centre for Medical Biotechnology, University Duisburg-Essen, Germany* [10] *Translational Neuro-Oncology, West German Cancer Center, University Hospital Essen, University Duisburg-Essen, Germany* [11] *German Cancer Consortium (DKTK), Germany* [12] *German Cancer Research Center (DKFZ), Heidelberg, Germany*

[*] sangkyun.lee@tu-dortmund.de

Identifying relevant signatures for predicting clinical outcome is a fundamental task in high-throughput studies. However, reported signatures in studies, composed of features e.g. mRNAs, miRNAs, or SNPs, are often non-overlapping, even though they have been identified from similar experiments of the same type of disease. The lack of a consensus is mostly due to high-dimensionality in data, i.e. the number of candidate features is much larger than the sample size. In such cases, signature selection suffers from large variation.

We propose a robust signature selection method that enhances the selection stability of penalized regression algorithms, which identifies the best survival risk predictor in a reduced dimension. Our method is based on an aggregation of multiple (possibly unstable) signatures obtained with the preconditioned lasso [PBHT08] algorithm applied to random (internal) subsamples of a given cohort data, where the aggregated signature is shrunken by a simple thresholding strategy. The resulting method, RS-PL, is conceptually simple and easy to apply, relying on parameters automatically tuned by cross validation. Robust signature selection using RS-PL operates within an (external) subsampling framework to estimate the selection probabilities of features in multiple trials of RS-PL. These probabilities are used for identifying stable features, and thereby for building a robust signature.

Our method was evaluated on microarray data sets from neuroblastoma, lung adenocarcinoma, and breast cancer patients, extracting robust and relevant signatures for predicting survival risk. Signatures obtained by our method achieved high prediction performance and robustness, consistently over the three data sets. Genes with high selection probability in our robust signatures have been reported as cancer-relevant. The software is available as an R package `rsig` at CRAN (`http://cran.r-project.org`), and the full paper has been published in PLoS ONE [LRL+14].

References

[LRL+14] Sangkyun Lee, Jörg Rahnenführer, Michel Lang, Katleen De Preter, Pieter Mestdagh, Jan Koster, Rogier Versteeg, Raymond L. Stallings, Luigi Varesio, Shahab Asgharzadeh, Johannes H. Schulte, Kathrin Fielitz, Melanie Schwermer, Katharina Morik, and Alexander Schramm. Robust selection of cancer survival signatures from high-throughput genomic data using two-fold subsampling. *PLoS ONE*, 9(10):e108818, 2014.

[PBHT08] Debashis Paul, Eric Bair, Trevor Hastie, and Robert Tibshirani. "Preconditioning" for feature selection and regression in high-dimensional problems. *Annals of Statistics*, 36(4):1595–1618, 2008.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Functional Analysis of Metabolic Networks using Sparse Group Lasso

Satya Swarup Samal[1], Andreas Weber[2] and Holger Fröhlich[1]
(1) Algorithmic Bioinformatics, Institute for Computer Science, University of Bonn, c/o Bonn-Aachen International Center for IT, D-53113, Bonn, Germany
(2) Institut für Informatik II, University of Bonn, Friedrich-Ebert-Allee 144, 53113 Bonn, Germany
(1) samal@cs.uni-bonn.de (2) weber@cs.uni-bonn.de (3) frohlich@bit.uni-bonn.de

Integration of metabolic networks with "omics" data (particularly microarray based gene expression data) has been a subject of recent research [RPT+15]. Under the assumption of steady state of underlying biochemical system, metabolic networks can be algebraically decomposed into a set of pathways, which are simplest steady state flux distributions. In literature, such pathways are sometimes referred to as extreme currents (ECs), elementary flux modes (EFMs) or extreme pathways (EPs), depending on the way, in which reversible reactions are split [LP10]. Enzymes associated to active reactions (with nonzero flux) can be mapped to genes, resulting in a gene set for each pathway. The question then is to compute the associations of such a gene set with a given phenotype or clinical outcome. This may be understood as a particular instance of a self contained gene set test [GvdGdKvH04]. In this direction, we propose a method based on sparse group lasso (SGL) [SFHT13] to identify phenotype associated ECs based on gene expression data. SGL selects a sparse set of feature groups and also introduces sparsity within each group. Features in our model are clusters of ECs, and feature groups are defined based on correlations among these features [BRvdGZ13]. We apply our method to a list of metabolic networks from KEGG database, compute ECs and study the association of EC clusters to clinical phenotypes in prostate cancer (where the outcome is tumor and normal, respectively) as well as glioblastoma multiforme (where the outcome is survival). We provide simulations to show the superior performance of our method compared to the global test [GvdGdKvH04]. The advantage of our approach is two-fold. First, high gene set level overlap between ECs is addressed by representing the features as "clusters of EC" and correlations among features is addressed by defining groups, thereby addressing the non-identifiability of individual features. Second, the approach is flexible to analyse different types of clinical outcomes e.g. categorical (cancer vs healthy) or real valued (survival times). The advantage of using EC over EFM is to provide the link to well established techniques of stoichiometric network analysis e.g. computing hopf bifurcation [EEG+15] which are formalised using EC.

References

[BRvdGZ13] Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835 – 1858, 2013.

[EEG+15] Hassan Errami, Markus Eiswirth, Dima Grigoriev, Werner M. Seiler, Thomas Sturm, and Andreas Weber. Detection of Hopf bifurcations in chemical reaction networks using convex coordinates. *Journal of Computational Physics*, 291:279–302, 2015.

[GvdGdKvH04] Jelle J. Goeman, Sara A. van de Geer, Floor de Kort, and Hans C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[LP10] Francisco Llaneras and Jesús Picó. Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 753904, 13 pages, 2010.

[RPT+15] Alberto Rezola, Jon Pey, Luis Tobalina, ngel Rubio, John E. Beasley, and Francisco J. Planes. Advances in network-based metabolic pathway analysis and gene expression data integration. *Briefings in Bioinformatics*, 16(2):265–279, 2015.

[SFHT13] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

EosinophilDetector - A new approach for recognizing Eosinophiles within label-free CARS images

Thomas Temme
Department of Biophysics, Ruhr-University Bochum
thomas.temme@bph.rub.de

Eosinophiles as white blood cells are an essential component of the immune system, and as such are known to play an important role in both colon cancer and colitis. Early studies have shown that the number of eosinophils found in colon cancer tissue correlates with the survival rate of the patients, so that quantifying their number in a given tissue section is of high diagnostic relevance. We have established a label-free CARS approach which provides a non-invasive and fast way for scanning relevant tissue section and allows to identify and count eosinophils.

In this approach, identification and counting of eosinophiles requires substantial computational efforts. In order to achieve reliable and robust results our computational analysis follows a mostly unsupervised strategy. Thus, our approach starts with clustering the original CARS images. As it turns out, spectral patterns associated with eosinophils are distinctive from other parts of the tissue, so that they can be identified with clusters in clustering-based segmentations. Reliable recognition and counting of eosinophils is reduced to identifying the eosinophil clusters in an automated fashion. We accomplish this by applying different morphological filters on the cluster images. Therefore every connected component within each cluster is measured for certain morphological features like size and roundness. Every cluster which has a certain number of connected components fitting to that predefined morphological attributes is treated as eosinophil cluster. Finally all recognized eosinophil-clusters are treated as one big cluster and are morphological postprocessed to reduce noise and artifacts. Every remaining connected component in that big cluster is counted as an eosinophil.

We validated our approach by comparing to eosinophils counted manually by a pathologist. Our results demonstrate that our novel algorithm is able to reliably estimate the amount of eosinophils on label-free CARS data.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Discovering Associations between Patient Phenotypes and Gene Functions

Lara Urban[1,2,*], Christian W Remmele[1], Roland F Schwarz[3], Pietro Liò[2], Marcus Dittrich[1,4], and Tobias Müller[1,*]

[1]Department of Bioinformatics, Biocenter, Julius-Maximilians-University of Würzburg, Am Hubland, 97074 Würzburg, Germany
[2]Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge, CB3 0FD, United Kingdom
[3]European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
[4]Department of Human Genetics, Biocenter, Julius-Maximilians-University of Würzburg, Am Hubland, 97074 Würzburg, Germany
[*]lara.h.urban@gmail.com, tobias.mueller@biozentrum.uni-wuerzburg.de

With the current abundance of gene expression data, a challenge for transcriptomic analysis is the extraction of biological meaning. To supplement established functional analyses such as Gene Ontology Enrichment and Gene Set Enrichment Analysis, we assessed and adapted multivariate ecological methods for transcriptomic data. RLQ ordination and fourth-corner analysis are widely used in ecology to discover associations between species traits and environmental variables. The RLQ ordination visualizes associations within and between covariates by spatial proximity, while the fourth-corner analysis identifies the significance of the associations.

Here, we used these methods to discover associations between patient phenotypes and gene functions. Application to gene expression data of patients suffering from acute lymphocytic leukemia identified multiple significant associations between patient phenotypes and gene functions. For instance, T-cell specific gene functions were found to be positively associated with T-cell origin of the cancer cells and negatively with B-cell origin.

For ensuring optimal performance of these methods during analysis of transcriptomic data, various data transformation procedures were introduced and assessed. Power and specificity of the fourth-corner analysis were validated. In summary, our study demonstrated the high potential of RLQ and fourth-corner analysis to unravel complex gene-phenotype associations in large scale transcriptomic datasets.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Mobile elements present a challenge in the analysis of Bacteroides vulgatus mpk

Sina Beier[1], Anna Lange[2], Daniel H. Huson[1], Ingo B. Autenrieth[2] and Julia-Stefanie Frick[2]
[1] *ZBIT Center for Bioinformatics, University of Tübingen*
[2] *Interfacultary Institute for Microbiology and Infection Medicine, University of Tübingen*
sina.beier@uni-tuebingen.de

We have sequenced, assembled and investigated the genome of *Bacteroides vulgatus* mpk, a bacterium isolated from the gut of healthy mice which has been shown to prevent *Escherichia coli*-induced colitis in mice and is thus a potential probiotic [WBF$^+$03]. Though *B. vulgatus* is commonly found in murine and human gut microbiota there are only few findings on genome composition and horizontal gene transfer (HGT) [XML$^+$07]. In fact, there is only one complete reference genome for this species, as well as multiple highly fragmented draft assemblies.

We initially suspected the high fragmentation of all available short read assemblies of this species to be caused by the large amount of mobile elements known to be found in *Bacteroides* genomes, which increase assembly complexity [KSP10]. We therefore decided to use long read sequencing to be able to produce a draft genome of *B. vulgatus* mpk which is a sufficient basis for further analysis, including comparative genomics. We achieved this goal using an assembly, annotation and analysis pipeline tailored to the needs of a bacterial genome including a large amount of paralogy and HGT.

Investigation of our draft genome shows that *B. vulgatus* mpk in fact harbours a large number of mobile elements including transposable elements, conjugative transposons and a CRISPR/Cas system inserted in the genome. Comparative genomics showed that most of the sequences unique to *B. vulgatus* mpk in comparison to the typestrain *B. vulgatus* ATCC 8482 and other *Bacteroides* are comprised of mobile elements and that *B. vulgatus* mpk includes mobile elements in comparatively high copy numbers and conservation. This could also be an artefact of the short read assemblies of other strains, which might have wrongly reduced the number of paralogs due to the challenges of handling repeated sequence.

The mobilome of *B. vulgatus* is thus the most important factor driving the variability of this species, but also the factor making genome assembly and correct annotation a much harder task and the information available much less than it would be expected for one of the most common bacteria in the mammal gut.

References

[KSP10] Carl Kingsford, Michael C Schatz, and Mihai Pop. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics*, 11:21, 2010.

[WBF$^+$03] Marc Waidmann, Oliver Bechtold, Julia-Stefanie Frick, Hans-Anton Lehr, Sören Schubert, Ulrich Dobrindt, Jürgen Loeffler, Erwin Bohn, and Ingo B Autenrieth. Bacteroides vulgatus protects against Escherichia coli-induced colitis in gnotobiotic interleukin-2-deficient mice. *Gastroenterology*, 125(1):162–177, July 2003.

[XML$^+$07] Jian Xu, Michael Mahowald, Ruth E Ley, Catherine Lozupone, Micah Hamady, Eric C Martens, Bernard Henrissat, Pedro M Coutinho, Patrick Minx, Philippe Latreille, Holland Cordum, Andrew Van Brunt, Kyung Kim, Robert S Fulton, Lucinda Fulton, Sandra W Clifton, Richard K Wilson, Robin D Knight, and Jeffrey I Gordon. Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biology*, 5(7):e156, July 2007.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

N-gram analysis of amyloid data

Michał Burdukiewicz[1], Piotr Sobczyk[2], Paweł Mackiewicz[1] and Małgorzata Kotulska[3]
[1] *University of Wrocław, Department of Genomics, Poland*
[2] *Wrocław University of Technology, Department of Mathematics, Poland*
[3] *Wrocław University of Technology, Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Poland*
malgorzata.kotulska@pwr.edu.pl

Amyloids are short proteins associated with a number of clinical disorders, for example Alzheimer's or Creutzfeldt-Jakob's diseases. Despite their variability in size and amino acid composition, most amyloidogenic sequences form cytotoxic aggregates, although a few aggregates are biologically functional [BU15]. The hallmark trait of amyloids is the presence of characteristic short sequences of amino acids, called hot-spots. Furthermore, amyloids can create zipper-like β-structures [F12]. Although studies investigating properties of amyloidogenic sequences have already been conducted, the newly established AmyLoad database facilities large-scale analysis of amyloids [WK15].

Among commonly acclaimed methods of predicting amyloids, FISH Amyloid [GK14] focuses more on the putative motifs of hot spots. To expand its model by considering longer and more complicated motifs, we used n-gram analysis. N-grams (k-mers) are vectors of n characters derived from input sequences. The number of possible n-grams is equal to n^u, where u is the length of the alphabet (4 in case of nucleic acids and 20 in case of proteins). To deal with the dimensionality of the problem, we implemented QuiPT (Quick Permutation Test) in *biogram* software [BSL15], which performs an exact test instead of a large number of permutations.

To reduce the dimension even more, we grouped amino acids into clusters based on their physicochemical properties potentially important in the amyloid type of aggregation. The features include several scales quantitatively representing hydrophobicity, size and accessibility derived from AAIndex database, and propensity of amino acids to form contact sites derived in [WK14].

The n-gram model, trained on the data from AmyLoad database, is validated through amyloid prediction framework using random forests. The preliminary analysis of the amyloidogenic sequences not only facilitates prediction of amyloids but also gives a new insight into the physicochemical characteristics of the hot spots. The mean AUC of the classifier committee in 5-fold cross-validation was 0.89. The most balanced classifier, regarding its sensitivity S_n and specificity S_p, enabled the predictions with $S_n = 0.75$, $S_p = 0.87$, and AUC $= 0.89$.

The address of amylogenicity predictor (AmyloGram): www.smorfland.uni.wroc.pl/amylogram.

References

[BSL15] Michal Burdukiewicz, Piotr Sobczyk, and Chris Lauber. *biogram: analysis of biological sequences using n-grams.* 2015. R package version 1.2.

[BU15] Leonid Breydo and Vladimir N. Uversky. Structural, morphological, and functional diversity of amyloid oligomers. *FEBS letters*, July 2015.

[F12] Marcus Fndrich. Oligomeric Intermediates in Amyloid Formation: Structure Determination and Mechanisms of Toxicity. *Journal of Molecular Biology*, 421(45):427–440, August 2012.

[GK14] Pawel Gasior and Malgorzata Kotulska. FISH Amyloid a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC Bioinformatics*, 15(1):54, February 2014.

[WK14] Pawel P. Wozniak and Malgorzata Kotulska. Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling*, 20(11), 2014.

[WK15] Pawel P. Wozniak and Malgorzata Kotulska. AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics (Oxford, England)*, June 2015.

Generation of artificial count data for the simulation of feature subsets in NGS or proteomics experiments

Jochen Kruppa, Frank Kramer, Tim Beißbarth and Klaus Jung
Department of Medical Statistics, University Medical Center Göttingen
klaus.jung@ams.med.uni-goettingen.de

Count data occur frequently in next-generation sequencing experiments, for example when the number of reads mapped to a particular region of a reference genome is counted [LSS14]. Count data are also typical for proteomics experiments making use of affinity purification combined with mass spectrometry, where the number of peptide spectra that belong to a certain protein is counted [CFN08]. Therefore, the simulation of correlated count data is important for validating new statistical methods for the analysis of the mentioned experiments. Nonetheless, the current methods for simulating count data from sequencing experiments only consider single runs with uncorrelated features, while the correlation structure is explicitly of interest for several statistical methods.

As one possible solution, we propose to draw correlated data from the multivariate normal distribution and to round these continuous data in order to obtain discrete counts. In our approach, the required distribution parameters can either be constructed or estimated from real count data. Rounding might affect the correlation structure. Therefore, we evaluate the use of shrinkage estimators that have already been used in the context of microarray experiments [SS05, LW03]. With our approach it is not possible to generate correlated data for the features of a whole experiment, however, our approach turned out to be useful for the simulation of counts for defined subsets of features like individual pathways or GO categories.

In a simulation study we found that there are less deviations between the covariance matrices of rounded and unrounded data when using the shrinkage estimators proposed by Schäfer and Strimmer (2005) [SS05]. We demonstrate that the methods are useful to generate artificial count data for certain predefined covariance structures (e.g. autocorrelated, unstructured) but also for covariance structures estimated from real data examples. We demonstrate the applicability on a public available data set from AffyExpress.

References

[CFN08] H. Choi, D. Fermin, and A.I. Nesvizhskii. Significance Analysis of Spectral Count Data in Label-free Shotgun Proteomics. *Moll Cell Proteomics*, 7:2373–2385, 2008.

[LSS14] Y. Liao, G.K. Smyth, and W. Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30:923–930, 2014.

[LW03] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J Empir Fianc*, 10:603–621, 2003.

[SS05] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance estimation and implications for functional genomics. *Statist Appl Genet Mol Biol*, 4, 2005.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Identification, expression and alternative splicing analysis of Ribosome biogenesis factors in Tomato

Stefan Simm[1,2], Mario Keller[1], Sotirios Fragkostefanakis[1,2], Enrico Schleiff[1,2,3,4]
[1] *Goethe University,Department of Biosciences, Molecular Cell Biology of Plants;*[2] *Cluster of Excellence Frankfurt;*[3] *Center of Membrane Proteomics;*[4] *Buchmann Institute of Molecular Life Sciences; Max von Laue Str. 9, 60438 Frankfurt/Main, Germany*
ssimm@bio.uni-frankfurt.de

Ribosome biogenesis involves a large inventory of proteinaceous and RNA cofactors. More than 250 ribosome biogenesis factors (RBFs) have been described in yeast [ESL+14]. However, information on plant ribosome biogenesis in general is rather sparse and functional relevance has not been experimentally approached yet. An overlap of nearly 75% between RBF inventory in yeast and plants could be observed. Nevertheless, the complexitiy of orthologous groups in single plant species was not analyzed, which is of interest to explore tissue, developmental and condition specific RBFs.

These Ribosome biogenesis factors are involved in multiple aspects like rRNA processing, folding and modification as well as in ribosomal protein (RP) assembly. Considering the importance of RBFs for particular developmental processes, we examined the complexity of RBF and RP (co-)orthologs by bioinformatic assignment in 14 different plant species and expression profiling in the model crop Solanum lycopersicum.

Assigning (co-)orthologs to each RBF revealed that at least 25% of all predicted RBFs are encoded by more than one gene. At first we realized that the occurrence of multiple RBF co-orthologs is not globally correlated to the existence of multiple RP co-orthologs [SFP+15].

The transcript abundance of genes coding for predicted RBFs and RPs in leaves and anthers of *S. lycopersicum* was determined by next generation sequencing (NGS). In combination with existing expression profiles, we could conclude that co-orthologs of RBFs by large account for a preferential function in different tissue or at distinct developmental stages. This notion is supported by the differential expression of selected RBFs during male gametophyte development [MWM+13]. In addition, co-regulated clusters of RBF and RP coding genes have been observed.

Beside the different expression pattern of co-orthologous RBFs and RPs we analyzed further the presence of different splicing isoforms in the reproductive tissue pollen. The existence of alternative splicing events like intron retentions and exon skippings in specific tissues gives insight in genetic regulation [TZL+14].

References

[ESL+14] Ingo Ebersberger, Stefan Simm, Matthias S. Leisegang, Peter Schmitzberger, Arndt von Haeseler, Markus T. Bohnsack, and Enrico Schleiff. The evolution of the ribosome biogenesis pathway from a yeast perspective. *Nucleic Acids Research*, 42:1509–1523, 2014.

[MWM+13] Sandra Missbach, Benjamin L. Weis, Roman Martin, Stefan Simm, Markus T. Bohnsack, and Enrico Schleiff. 40S ribosome biogenesis co-factors are essential for gametophyte and embryo development. *PLoS One*, 8:e54084, 2013.

[SFP+15] Stefan Simm, Sortiris Fragkostefanakis, Puneet Paul, Mario Keller, Jens Einloft, Klaus-Dieter Scharf, and Enrico Schleiff. Identification and Expression Analysis of Ribosome Biogenesis Factor Co-orthologs in Solanum lycopersicum. *Bioinformatics and Biological Insights*, 9:1–17, 2015.

[TZL+14] Shawn R. Thatcher, Wengang Zhou, April Leonard, Bing-Bing Wang, Mary Beatty, Gina Zastrow-Hayes, Xiangyu Zhao, Andy Baumgarten, and Bailin Li. Genome-Iwde Analysis of Alternative Splicing in *Zea mays*: Landscape and Genetic Regulation. *The Plant Cell*, 26:3472–3487, 2014.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Simultaneous Gene Finding in Aligned Genomes

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

With recent technologies in whole-genome sequencing, the sequencing of entire clades of related genomes is in progress. The Genome 10K project, for example, aims at sequencing 10 000 vertebrate species. Other examples include the 5 000 Insect Genome Project (i5k) and the 1 000 Fungal Genomes Project of the JGI. In these and other sequencing efforts subclades are sequenced whose member genomes are close enough to have widely conserved gene structures, yet diverse enough so information from mutation and selection can be exploited. Comparative gene finding approaches, such as N-Scan [GB06] and Contrast [GDSB07] have been developed to leverage multiple genome alignments, e.g. exploiting conservation patterns.

We have extended the gene-finder Augustus [SKG+06] for comparative gene prediction. The new Augustus-cgp takes a whole-genome alignment of multiple closely related species and simultaneously predicts the protein-coding genes in *all* input genomes. This approach is conceptually different from previous programs which model only the structure of a single (N-Scan, Contrast) or two genomes (pair-HMMs) at a time.

The identification of genes in multiple genomes can be stated as a discrete optimization problem on a graph. Nodes in the graph represent candidate exons. For each species, a path from a source to a sink node represents candidate gene structures. The score of a joint gene structure, that is exactly one path for each species, takes into account both intrinsic evidence from content and signal models and extrinsic evidence such as transcriptome data (e.g. RNA-Seq) or existing annotations. Furthermore, it includes a cross-species score which considers the phylogeny, conservation of sequence and exon boundaries, selective pressure, etc. Although maximizing the score of a joint gene structure is NP-complete, in many cases an exact solution can still be found using a dual decomposition approximation.

We applied Augustus-cgp to predict the genes in 12 *Drosophila* genomes aligned with Cactus [PEN+11] and evaluated it on *D. melanogaster*. The comparative approach improves the accuracy of Augustus single-species gene finding both when no extrinsic evidence is used (*ab initio*) and when RNA-Seq data is included for a subset of the input genomes. In the *de novo* category, where only the aligned genomes are input, Augustus-cgp outperforms N-Scan. The task of exploiting a trusted annotation of a related species for the annotation of a new target genome, can be considered a special case of comparative gene-finding. Augustus-cgp outperforms the homology-based gene-finder GeneWise [BCD04] for transferring the annotation of *D. simulans* to *D. melanogaster*.

References

[BCD04] E. Birney, M. Clamp, and R. Durbin. GeneWise and Genomewise. *Genome Res.*, 14:988–995, 2004.

[GB06] Samuel S. Gross and Michael R. Brent. Using Multiple Alignments to Improve Gene Prediction. *J. Comp. Biol.*, 13(2):379–393, 2006.

[GDSB07] Samuel S. Gross, Chuong B. Do, Marina Sirota, and Serafim Batzoglou. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol.*, 8(12):R269+, 2007.

[PEN+11] B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, and D. Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.*, 21(9):1512–1528, 2011.

[SKG+06] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, 34:435–439, 2006.

Poster 19

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

A Generic Approach to Antigen Design Based on Minimising Sequence Distances

Jan T Kim, Tom Peacock, Munir Iqbal
The Pirbright Institute, Ash Road, Pirbright, Woking GU24 0NF, UK
jan.kim@pirbright.ac.uk

Protection against a broad spectrum of genetic variants is one of the key objectives in designing vaccines against quickly evolving RNA viruses. While cross-protection depends on many complex factors (including 3D molecular structures, immune system processes, and others), sequence similarity is, overall, strongly correlated to cross-reactivity to antisera. Therefore, minimising sequence distance (or equivalently, maximising similarity) can be used as a criterion to design a broadly reactive antigen.

The consensus of a set of sequences minimises distance and is thus a point of departure for antigen design. However, if some antigen clades are over-represented in the set, this biases the consensus so that the distance to the other clades becomes larger than necessary. Domain specific knowledge can be used to counter such biases, e.g. intermediary consensus sequences at outbreak group and subclade levels of H5N1 influenza virus strains have been used for this purpose [GR11]. This approach is equivalent to placing variable weights on groups of sequences in order to compensate for their over- or under-representation.

In continuous sequence space [VS93], the mean of a set of sequences corresponds to the nucleotide frequency matrix which is typically used to represent sequence motifs [Sto00] such as transcription factor binding sites (TFBS). Specificity of such scoring matrices can be improved by weighting and it is maximised by the Binding Matrix (BM) [KGM04]. The BM is related to classification of samples from subgaussian distributions [MMKB+03] and concentrates weight on the most distant sequences. This is relevant to vaccine design because achieving cross-reactivity with distant antigens is critical while further optimisation is unnecessary once cross-reactivity is sufficient for protection. We have adapted the BM / subgaussian approach for computing a consensus that minimises maximal distance, and demonstrate its application to a set of H9N2 haemagglutinin sequences.

References

[GR11] Brendan M. Giles and Ted M. Ross. A Computationally Optimized Broadly Reactive Vaccine Elicits Broadly Reactive Antibodies in Mice and Ferrets. *Vaccine*, 29:3043–3054, 2011.

[KGM04] Jan T. Kim, Jan E. Gewehr, and Thomas Martinetz. Binding Matrix: A Novel Approach for Binding Site Recognition. *Journal of Bioinformatics and Computational Biology*, 2:289–307, 2004.

[MMKB+03] Amir Madany Mamlouk, Jan T. Kim, Erhardt Barth, Michael Brauckmann, and Thomas Martinetz. One-Class Classification with Subgaussians. In Bernd Michaelis and Gerald Krell, editors, *DAGM Symposium*, pages 346–353, Berlin Heidelberg, 2003. Springer Verlag.

[Sto00] Gary D. Stormo. DNA Binding Sites: Representation and Discovery. *Bioinformatics*, 16:16–23, 2000.

[VS93] Martin Vingron and Peter R. Sibbald. Weighting in Sequence Space: A Comparison of Methods in Terms of Generalized Sequences. *Proceedings of the National Academy of Sciences, USA*, 90:8777–8781, 1993.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

The composition of the cyanobacterial core- and pan-genome

Stefan Simm[1,2], Mario Keller[1], Mario Selymesi[1], Enrico Schleiff[1,2,3,4]
[1] *Goethe University, Department of Biosciences, Molecular Cell Biology of Plants;*[2] *Cluster of Excellence Frankfurt;*[3] *Center of Membrane Proteomics;*[4] *Buchmann Institute of Molecular Life Sciences;*
Max von Laue Str. 9, 60438 Frankfurt/Main, Germany
ssimm@bio.uni-frankfurt.de

Cyanobacteria are photosynthetic procaryotes living in a huge variety of ecosystems and have high potential for biotechnological usage [GPJYB03]. In part, the understanding of functionality of cyanobacteria and their interaction with the environment can be inferred from the analysis of their genomes / proteomes. Meta-analysis, like pan-genome determination, describes the entirety of gene sets including all strains of a species or a set of species from a phylum. A pan-genome includes a core-genome, a dispensable-genome as well as unique genes [RHF+09]. The core-genome contains gene sets found in all analyzed species / strains in contrast to the dispensable-genome, where only some species/strains contain the gene but at least more than one.

Pan-genome analysis can be used to identify biological pathways present in all cyanobacteria as proteins involved in such processes are encoded by the core-genome. However, beside identification of fundamental processes, genes specific for certain cyanobacterial features can be identified as well by analyzing the dispensable-genome. By this, we identified 559 clusters of likely orthologous groups (CLOGs) as core-genome of 58 analyzed cyanobacteria, whereof 20% are invovled in protein homeostasis and most of the other groups are involved in housekeeping function.

Analyzing the pan-genome of cyanobacteria 15,742 CLOGs were found in the dispensable genome at which most of them were found in less than 10 different cyanobacteria. In addition, it could be shown that the pan-genome is open and has an increase factor of 0.35. Furthermore, 3 (57) genes are likely to be signature genes for thermophilic (heterocyst-forming) cyanobacteria, respectively [SKSS15].

To get insights into cyanobacterial systems for the interaction with the environment, we also inspected the diversity of the outer membrane proteome with focus on β-barrel proteins having a great influence in nutrient import [MSN+09]. 919 protein sequences in all 58 cyanobacteria were predicted to be putative β-barrel proteins. Those β-barrel proteins could be assigned to 21 different clusters each containing more than 4 sequences. The different clusters were representing 12 functional groups of domains.

Most of the transporting outer membrane β-barrel proteins are not globally conserved and occurrence of β-barrel proteins shows high strain specificity. The core set of outer membrane proteins comprises three proteins only, namely Omp85, LptD and an OprB-type porin. Thus, we conclude that cyanobacteria have developed individual strategies for the interaction with the environment.

References

[GPJYB03] Ferran Garcia-Pichel, Susan L. Johnson, David Youngkin, and Jayne Belnap. Small-scale vertical distribution of bacterial biomass and diversity in biological soil crusts from arid lands in the Colorado plateau. *Microbial Ecology*, 46:312–321, 2003.

[MSN+09] Oliver Mirus, Sascha Strauss, Kerstin Nicolaisen, Arndt von Haeseler, and Enrico Schleiff. TonB-dependent transporters and their occurrence in cyanobacteria. *BMC Biology*, 7:68, 2009.

[RHF+09] Michael L. Reno, Nicole L. Held, Christopher J. Fields, Patricia V. Burke, and Rachel J. Whitaker. Biogeography of the Sulfolobus islandicus pan-genome. *Proc Natl Acad Sci U S A*, 106(21):8605–8610, 2009.

[SKSS15] Stefan Simm, Mario Keller, Mario Selymesi, and Enrico Schleiff. The composition of the global and feature specific cyanobacterial core-genomes. *Frontiers in Microbiology*, 6:219, 2015.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

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

Annett Erkes[1], Jan Grau[1], Maik Reschke[2], Jana Streubel[2], Richard D. Morgan[3], Geoffrey G. Wilson[3],
Ralf Koebnik[4], Jens Boch[2]
[1] *Institute of Computer Science, Martin Luther University Halle–Wittenberg*
[2] *Department of Genetics, Martin Luther University Halle–Wittenberg*
[3] *New England Biolabs Inc., Ipswich, MA 01938, USA.*
[4] *UMR 186 IRD-UM2-Cirad "Résistance des Plantes aux Bioagresseurs", BP 64501, 34394 Montpellier cedex 5, France.*
annett.erkes@informatik.uni-halle.de

Plant-pathogenic *Xanthomonas* bacteria use transcription activator-like effectors (TALEs) that bind to the promoter of plant genes and activate their transcription. *Xanthomonas* infections result in a substantial yield loss for many crop plants including rice. The binding domain of TALEs consists of tandem repeats containing two hypervariable amino acids, which are called repeat variable di-residue (RVD). Each RVD recognizes one nucleotide of its target DNA and the consecutive array of RVDs determines TALE target specificity.

Here, we present AnnoTALE, an application for annotating TALEs in *Xanthomonas* genomes, for analyzing their structure, for clustering TALEs by the similarity of their RVD sequences, and for predicting putative TALE target genes. We sequence the genome of *Xanthomonas oryzae pv. oryzae* PXO83 by PacBio sequencing and use AnnoTALE to predict all TALE genes of this strain. Building classes of TALEs from published and the newly sequenced *Xanthomonas* genomes allows us to gain new insights into TALE evolution.

We find that RVDs are highly conserved even on the codon level. We discover that only one to three codon pairs coding for one RVD occur in known TALEs, even though the number of theoretically possible codon pairs is substantially larger. In addition, some codon pairs found frequently in RVDs are rather rare in the remaining TALE sequence and in coding sequences of other genes.

We compare the aligned RVDs of the class members and generate a network of possible and observed substitutions and find only one synonymous substitution between two codon pairs for RVD NN, whereas the remaining substitutions lead to a modification of the RVD. Most frequently, only one nucleotide is substituted between the compared RVDs of two class members. Our findings indicate that one way how TALE specificities evolve is by direct base substitutions in RVD codons.

Polar Plots for Visualizing the Effects of Different Treatments on Gene Regulation

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

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

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

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

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

To asses whether the directional pattern of gene responses significantly differs from a uniform distribution, we apply Watson's goodness of fit test [Ste70, JS01], and for comparing two different gradient distributions, e.g. for two different herbivores, we use Watson's two-sample test of homogenity [JS01].

References

[JS01] S Rao Jammalamadaka and Ambar Sengupta. *Topics in circular statistics*, volume 5. World Scientific, 2001.

[Ste70] M. A. Stephens. Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(1):pp. 115–122, 1970.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Motif independent recognition of microRNA precursor sequences

Thorsten Schmidt, Thilo Wilts
Department of Natural Sciences, Hochschule Emden Leer
thorsten.schmidt@hs-emden-leer.de

MicroRNAs (miRNA) are a class of short non-coding RNA sequences which play an important role in various regulation processes and can have significant influence in certain cancer types and other diseases like Alzheimer[SRS07]. During biogenesis the mature microRNA are -among other steps- processed from microRNA precursors which usually show a typical stem-loop secondary structure and well conserved sequences [SRS07]. Thus, the current identification of microRNA precursors is based on a combination of a high number of typical sequence motifs and structure properties mainly [LSC14, BP09, JWW^{+}07]. This leads to high computational costs for large data sets e.g. from NGS experiments[LSC14]. Additionally, the application of a myriad of features leads to overfitting likely. Beyond memorization of sequence and structure patterns is of limited practical use to identify new, different microRNAs. Here, we present two novel features helping to identify microRNA precursors which are i) fast to calculate and ii) fundamentally different to known approaches.

As test sets we used the data used by the work of Lopes et al [LSC14] and the described evaluation methodology therein. Our first novel feature is the number of theoretically feasible secondary structures for a given nucleotide sequence. When calculating not only the typical hairpin like secondary structure with the lowest free folding energy but also structures with higher free folding energies, we found that miRNA precursor sequences tend to have less secondary structures (mean: 17.6) when compared to non-precursor sequences (mean: 47.1). In a biological context this seems to imply that precursor sequences tend to have a lower folding flexibility which can be a crucial characteristic recognized by further cellular processing steps. Our second novel feature is the structural complexity of the secondary structures. When transferring the dotbracket sequences of a given nucleotide sequence into their most abstract form [GVR04] where only the length-independent characteristically loop structure is remained, we found that miRNA precursor sequences tend to have a lower complexity in their secondary structures (mean: 1.9) compared to non-precursor sequences (mean: 2.7).

Using just these two novel features, a prediction accuracy of 86.5% can be achieved already. Further advantages are i) lower computational costs than the reported set of the best-performing 13 features by Lopes et al [LSC14], and ii) decreased danger of error-prone overfitting as with the use of more features like in previous approaches. Beyond our novel features may be more oriented on the actual biology than approaches which simply memorize sequence- or structural motifs. Finally our here reported features are essentially different from previous ones which were used to identify microRNA precursors.

References

[BP09] Rukshan Batuwita and Vasile Palade. microPred : effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, 25(8):989–995, 2009.

[GVR04] Robert Giegerich, Björn Voss, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic acids research*, 32(16):4843–51, January 2004.

[JWW^{+}07] Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun, and Zuhong Lu. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic acids research*, 35(Web Server issue):W339–44, July 2007.

[LSC14] Ivani De O N Lopes, Alexander Schliep, and André P De L F De Carvalho. The discriminant power of RNA features for pre-miRNA recognition. *BMC bioinformatics*, 15(1):124, May 2014.

[SRS07] Harris S Soifer, John J Rossi, and På l Saetrom. MicroRNAs in disease and potential therapeutic applications. *Molecular therapy : the journal of the American Society of Gene Therapy*, 15(12):2070–9, December 2007.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

APP is a functional regulator at the presynaptic active zone

Martin Wegner[1], Melanie Laßek[2], Jens Weingarten[2], Jörg Ackermann[1], Walter Volknandt[2], and Ina Koch[1]

[1] *Department of Molecular Bioinformatics, Johann Wolfgang Goethe-University Frankfurt/Main,*
[2] *Department of Molecular and Cellular Neurobiology and Neuroscience, Goethe-University Frankfurt/Main*
mawegner@stud.uni-frankfurt.de

The amyloid beta peptide (Aβ) is the main constituent of senile plaques in brains that are affected by Alzheimer's disease [GW84]. Aβ is one cleaving product of the amyloid precursor protein (APP), which has been allocated to the presynaptic active zone (PAZ), and identified as a constituent of the hippocampal PAZ proteome [LWE+13, LWAP+14]. The PAZ is a highly dynamic and flexible site of neurotransmitte release. The hippocampus is the central brain region involved in learning and memory and severely affected in the progression of AD. APP has been implicated in a variety of cellular functions, e.g., neuronal development, synapse formation, neurotransmitter release, and synaptic plasticity, which underpins its substantial role in cognitive functions. However, the precise physiological function of APP in the central nervous system system is still unknown.

We analyzed proteomic data derived from APP knockout (APP-KO) and conditional APP/APLP2 double knockout (NexCre-cDKO) mice to study the physiological function of APP [WLM+15]. The integration of the dataset and publicly available database knowledge into a protein-protein interaction (PPI) network resulted in a network of 615 proteins and 3 390 experimentally validated physical interactions. Protein abundance changes upon APP-KO and NexCre-cDKO were determined in mice (12 replicates) and mapped onto the PPI network of the hippocampal PAZ proteome. Topological analyses revealed a central role of APP at the PAZ. Functional modules were detected by applying community detection algorithms. Functional enrichment analyses identified specific up- and downregulated cellular functions upon APP deletion.

The conditional double knockout led to upregulation of proteins associated with energy metabolism, while downregulated proteins were mainly enriched in modules of structural organization. In the single knockout the distribution of up- and downregulated proteins was more homogeneous than in the double knockout. Our findings of alterations in the composition in both APP-mutants identified APP as a structural and functional regulator within the hippocampal PAZ network.

References

[GW84] George G Glenner and Caine W Wong. Alzheimer's disease and Down's syndrome: sharing of a unique cerebrovascular amyloid fibril protein. *Biochemical and Biophysical Research Communications*, 122(3):1131–1135, 1984.

[LWAP+14] Melanie Laßek, Jens Weingarten, Amparo Acker-Palmer, Sandra M Bajjalieh, Ulrike Müller, and Walter Volknandt. Amyloid precursor protein knockout diminishes synaptic vesicle proteins at the presynaptic active zone in mouse brain. *Current Alzheimer Research*, 11(10):971–980, 2014.

[LWE+13] Melanie Laßek, Jens Weingarten, Ulf Einsfelder, Peter Brendel, Ulrike Müller, and Walter Volknandt. Amyloid precursor proteins are constituents of the presynaptic active zone. *Journal of Neurochemistry*, 127(1):48–56, 2013.

[WLM+15] Jens Weingarten, Melanie Laßek, Benjamin F Müller, Marion Rohmer, Dominic Baeumlisberger, Benedikt Beckert, Jens Ade, Patricia Gogesch, Amparo Acker-Palmer, Michael Karas, et al. Regional Specializations of the PAZ Proteomes Derived from Mouse Hippocampus, Olfactory Bulb and Cerebellum. *Proteomes*, 3(2):74–88, 2015.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Assessing the genetic code optimality using multiobjective optimization approach

Małgorzata Wnętrzak*, Paweł Błażej, Paweł Mackiewicz
Faculty of Biotechnology, University of Wrocław, Wrocław, Poland
e-mail: malgorzata.wnetrzak@smorfland.uni.wroc.pl

A very popular theory about the orign of the standard genetic code postulates that it evolved to minimize effects of deleterious mutations and translational errors [FWK03]. This process involved most probably several incommensurable and often competing objectives. To study this subject, Santos and Monteagudo [SM11] considered possible adaptability of the genetic code based on evolutionary computation approach. However, they examined only single objective case. Subsequently, other authors [ddT15] studied this problem using multiobjective optimization technique. As objective functions, they used several types of costs of amino acid changes resulting from all possible single point mutations without distinction of the position in codon.

We also examined the problem of the genetic code optimality as a multiobjective optimization problem, however, in contrast to the previous works, we applied three objectives. We considered the costs of amino acid substitutions in three codon positions separately to asses the importance of particular codon positions. Thanks to that we did not need to ascribe arbitrary weights to three codon positions in the optimization process.

Our results show that under considered measures the canonical genetic code has a tendency to eliminate effects of harmful mutations because it is close to solutions which minimize values of the objective functions.

References

[ddT15] Lariza Laura de Oliveira, Paulo S L. de Oliveira, and Renato Tinós. A multiobjective approach to the genetic code adaptability problem. *BMC Bioinformatics*, 16:52, 2015.

[FWK03] Stephen J. Freeland, Tao Wu, and Nick Keulmann. The case for an error minimizing standard genetic code. *Orig Life Evol Biosph*, 33(4-5):457–477, Oct 2003.

[SM11] José Santos and Angel Monteagudo. Simulated evolution applied to study the genetic code optimality using a model of codon reassignments. *BMC Bioinformatics*, 12:56, 2011.

TRAPLINE: An Integrated Galaxy Pipeline for RNAseq Data Processing, Evaluation and Prediction

Markus Wolfien[1], Ulf Schmitz[2], Robert David[3] and Olaf Wolkenhauer[1]
1 *Department of Systems Biology and Bioinformatics, University of Rostock*
2 *Centenary Institute, University of Sydney*
3 *Reference and Translation Center for Cardiac Stem Cell Therapies, University of Rostock*
markus.wolfien@uni-rostock.de

Next Generation Sequencing (NGS) enables researchers to acquire deeper insights into cellular functions. The lack of standardised and automated methodologies poses a challenge for the analysis and interpretation of RNA sequencing data. We present a freely available, state-of-the-art bioinformatics workflow that integrates the best performing data processing, evaluation and predicting methods (bit.ly/TRAPLINE_RNAseq). Galaxy [BHJ14] is a free web-based platform for omics research that addresses the following needs: accessibility, reproducibility and transparency. Using these benefits we developed an easy to use, comprehensive, **T**ransparent, **R**eproducible and **A**utomated analysis **P**ipe**LINE** for RNAseq data processing and evaluation. TRAPLINE can be accessed via the public Galaxy page of TRAPLINE and is ready to use without restrictions or time consuming software installation.

Data Processing Modules: TRAPLINE guides researchers easily through a well annotated NGS data processing pipeline that includes FastQC, FASTQ Quality Trimmer and Clipper for pre-processing [BGVK+10, BHJ14], TopHat2 for genome alignment [TRG+12], Picard tools for multiple read correction which can be used for SNP analyses (http://broadinstitute.github.io/picard) and Cufflinks2, Cuffquant, Cuffdiff2 and CummeRbund for differential expression analysis and further visualization [TRG+12].

Data Evaluation and Prediction Modules: TRAPLINE annotates genes with the help of DAVID [HSL09], predicts spliced variants and enriched promoter sites [TRG+12] as well as miRNA targets [BWG+08] and protein-protein interactions [CABO+15] to enable users to obtain comprehensive insights of their analysed samples. Ultimately, a *.csv* file is build that includes all information which are ready to be used in ones favourite network tool (eg. Cytoscape, Cell Designer).

Using TRAPLINE, experimentalists will be able to analyse their data on their own without learning programming skills. Accessibility and sharing the data and results are worldwide possible. Furthermore, our developed data evaluation modules empower researchers to gain quickly in-depth insights into the biology underlying the investigated data. Our work for the first time introduces an automated and integrated Galaxy workflow including detailed data processing, evaluation and prediction modules.

References

[BGVK+10] D. Blankenberg, A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, et al. Manipulation of FASTQ data with Galaxy. *Bioinformatics*, 26(14):1783–5, 2010.

[BHJ14] D. Blankenberg and J. Hillman-Jackson. Analysis of next-generation sequencing data using Galaxy. *Methods Mol Biol*, 1150:21–43, 2014.

[BWG+08] D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander. The microRNA.org resource: targets and expression. *Nucleic Acids Res*, 36(Database issue):D149–53, 2008.

[CABO+15] A. Chatr-Aryamontri, B. J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res*, 43(Database issue):D470–8, 2015.

[HSL09] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[TRG+12] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3):562–78, 2012.

Can we predict environmental conditions from NGS biological observations?

Irina Maria Curuia[1], Daniel Hoffmann[1], Jens Boenigk[2], and Manfred Jensen[2]

[1]Bioinformatics Department , University of Duisburg-Essen
[2]Biodiversity Department , University of Duisburg Essen

Ecosystems are shaped by external environmental factors and by interactions between organisms. Here we ask how well we can predict environmental factors in fresh water ecosystems from community compositions. Specifically, we are interested in the compositions of microbial communities and what they tell about the respective environment. The practical implications of the study include applications in environmental monitoring and quality assessment. Knowledge of a model that predicts environmental factors from microbial community composition inferred from High-Throughput Sequencing (HTSeq) data would enable a new type of ecological monitoring, based on microbial presence/absence or abundance.

The current work flow of the computational analysis is as follows. The input is HTSeq data of small ribosomal subunit (SSU) amplicons. The SSU gene is a standard genomic "barcode" that is present in all known life forms, though in species-specific variants. The presence of a specific variant of this gene therefore allows to confirm the presence of the corresponding species at the sampling site. Moreover, the number of copies ("reads") of the gene observed in a HTSeq analysis is related to the abundance of that species in the sample. Various abundance cutoff-schemes were tested to possibly suppress noise from many low-abundance species. In addition to these genomic data, we use as input, quantities that describe the environmental conditions of the sample (temperature, pH, dissolved organic carbon, nutrients content, etc.).

First, the HTSeq data were transformed to make data from different samples possibly better comparable. Four different transformations were tested: (1) transformation to (binary) presence/absence data per species and sample, (2) relative abundance of each species in a sample, (3) Hellinger transformation of reads n_{ij} of species i at sampling site j to $\sqrt{n_{ij}/\sum_k n_{ik}}$, and finally (4) the identity transformation, ie the use of absolute read numbers.

Second, to limit the size of the model and to increase it's performance, we selected as predictor variables for the actual modeling a smaller number of species that more strongly correlated (Pearson correlation tests with p-values < 0.001) with environmental parameters.

Third, using the selected predictor variables, we trained random forest regression models to predict environmental parameters from the compositions of microbial communities. Finally, the correlation of predictions and experimental values was evaluated in a leave-one-out validation run.

In the process of developing this work flow, we found that using abundance cutoffs does not lead to consistent improvements. The transformation to presence/absence lead to the regression models with the best performance in terms of correlation of predicted and measured environmental parameters. In summary, we can state that for many environmental parameters a reasonable prediction is possible, though the quality can be sensitive to the specific work flow applied.

Poster 28

Bioinformatics Analysis of Alternative Splicing in the human-pathogenic fungus Candida glabrata

Patricia Sieber[1,2], Reinhard Guthke[2], Stefan Schuster[1] and Jörg Linde[2]
[1] *Department of Bioinformatics, Faculty of Biology and Pharmacy, Friedrich Schiller University Jena*
[2] *Research Group Systems Biology and Bioinformatics, Hans Knöll Institute, Leibniz Institute for Natural Product Research and Infection Biology, Jena*
patricia.sieber@uni-jena.de

The process of alternative splicing (AS) increases the functional complexity of an organism by differential post- and co-transcriptional processing. This mechanism is common in all eukaryotic species but still little is known about its regulation and effect in fungi [GSP$^+$14].
The presented work is aimed at a better understanding of the role of AS in human-pathogenic fungi, showing results of *Candida glabrata*. This fungal species is of clinical interest since it can infect the human gut and it is the second most common pathogenic *Candida* species. There are just a few known AS events in *C. glabrata* [LDW$^+$15].
RNA-Seq data were analyzed using five different AS detection tools and manual filters. It was possible to determine a number of AS events in *C. glabrata* under nitrosative stress, pH changes and addition of human neutrophils. Some detected differential spliced genes are present under multiple conditions and may have an important regulatory function in this fungus.
The identified genes will be experimentally validated. Further results can be used to gain new insights into the mechanism of AS, including multi-species comparison and examination of its influence on protein domains.

References

[GSP$^+$14] K. Gruetzmann, K. Szafranski, M. Pohl, K. Voigt, A. Petzold, and S. Schuster. Fungal Alternative Splicing is Associated with Multicellular Complexity and Virulence: A Genome-Wide Multi-Species Study. *DNA Res.*, 21:27–39, 2014.

[LDW$^+$15] J. Linde, S. Duggan, M. Weber, F. Horn, P. Sieber, D. Hellwig, K. Riege, M. Marz, R. Martin, R. Guthke, and O. Kurzai. Defining the transcriptomic landscape of Candida glabrata by RNA-Seq. *Nucleic Acids Res.*, 43:1392–1406, 2015.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

De-novo transcriptome assembly and the reconstruction of alternatively spliced isoforms

Lasse Feldhahn[1,2,3], Francois Buscot[1,3], Vsevolod Makeev[2,4,5,6], Mika Tarkka[1,3] and Ivo Grosse[1,2]
[1] *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig*
[2] *Institute of Computer Science, Martin-Luther University Halle-Wittenberg*
[3] *Department of Soil Ecology, UFZ - Helmholtz Centre for Environmental Research, Halle*
[4] *Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow*
[5] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow*
[6] *Moscow Institute of Physics and Technology, Dolgoprudny, Moscow*
lasse.feldhahn@informatik.uni-halle.de

De-novo transcriptome assembly of RNA-seq reads is one of the open challenges of bioinformatics. Several de-novo transcriptome assemblers exist, but none of them has been tailored to reconstruct alternatively spliced isoforms with high accuracy. Here, we present Asgad, a de-novo trancriptome assembler based on splicing graphs rather than De Bruijn graphs that uses information from whole reads rather than from overlappig kmers. We present preliminary results on simulated and real data that illustrate the de-novo assembly approach based on splicing graphs in comparison to existing de-novo assembly approaches based on De Bruijn graphs.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Poster 30

Graph-based analysis of CD30$^+$ cell distributions in Hodgkin Lymphoma

Hendrik Schäfer [1,*], Tim Schäfer [1,*] Jörg Ackermann [1], Norbert Dichter [1], Claudia Döring [2], Sylvia Hartmann [2], Martin-Leo Hansmann [2] and Ina Koch [1†]

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

[2]Dr. Senckenbergisches Institut für Pathologie, Universitätsklinikum Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

Hodgkin lymphoma (HL) is a type of B cell lymphoma and arises from germinal center B–cells [KRZ$^+$94]. The typical multi-nucleated cells associated with HL are called Hodgkin/Reed-Sternberg (HRS) cells. To diagnose the disease and identify the specific subtype, biopsies are taken, immunostained against CD30, and inspected under a microscope by pathologists. CD30 is a cell surface protein, and a marker for HRS cells. CD30 is also expressed by non-malignant, activated cells of the immune system. The microscope slides can be scanned to produce high-resolution digital whole slide images (WSI). We present an analysis of a database of WSIs which includes cases of both HL and lympadenitis (LA), an inflammation of the lymph node due to viral or bacterial infections. We determined the spatial distribution of CD30$^+$ cells in the lymph node tissue under malignant (HL) and reactive (LA) conditions.

CD30$^+$ cells were identified in the WSI using an extended version of our imaging pipeline [SSS$^+$13]. We defined *cell graphs* based on the positions and morphological properties of the immunostained cells, combining methods from digital image processing and network analysis. The cell graphs enable a deeper analysis of cell distributions within the WSI. We present an analysis of the vertex degree distribution of CD30 cell graphs and compare them to a suitable null model. It turns out that CD30 cell graphs show higher vertex degrees than expected by a random unit disk graph, suggesting clustering of the cells. We found that a gamma distribution is suitable to model the vertex degree distributions of CD30 cell graphs, meaning that they are not scale-free. Moreover, we compare the graphs for LA and two subtypes of HL. LA and HL showed different vertex degree distributions. The vertex degree distributions of the two HL subtypes NScHL and mixed cellularity HL (MXcHL) were similar.

Findings of this investigation provide objective parameters for CD30$^+$ cells in HL. The method can be extended for different immuno cell types, their shapes and localizations as well as their interactions. These data can be used for a deeper understanding of the biological meaning of interactions between CD30$^+$ tumor cells, and possibly also in the future to assess the impact of antitumor drugs, help with the prognosis and develop new therapies.

References

[KRZ$^+$94] R. Küppers, K. Rajewsky, M. Zhao, G. Simons, R. Laumann, R. Fischer, and M.-L. Hansmann. Hodgkin disease: Hodgkin and Reed-Sternberg cells picked from histological sections show clonal immunoglobin gene rearrangements and appear to be derived from B cells at various stages of development. *Proceedings of the National Academy of Sciences of the United States of America*, 91:10962–10966, 1994.

[SSS$^+$13] T. Schäfer, H. Schäfer, A. Schmitz, J. Ackermann, N. Dichter, C. Döring, S. Hartmann, M.-L. Hansmann, and I. Koch. Image database analysis of Hodgkin lymphoma. *Computational Biology and Chemistry*, 46:1–7, 2013.

[*] shared first authorship

[†] to whom correspondence should be addressed

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Bioinformatics Analysis of Heterogeneous Data Reveals Characteristic Mutational Landscapes of Neuroblastoma Relapses

Corinna Ernst[1], Johannes Köster[2], Daniela Beißer[1], Christopher Schröder[1],
Marc Schulte[3], Alexander Schramm[3] and Sven Rahmann[1]

[1] *Genome Informatics, Institute of Human Genetics,*
University Hospital Essen, University of Duisburg-Essen, Essen, Germany

[2] *Center for Functional Cancer Epigenetics, Departments of Medical Oncology and Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA*

[3] *Pediatric Oncology and Hematology,*
University Children's Hospital Essen, University of Duisburg-Essen, Essen, Germany

corinna.ernst@uni-due.de

We comprehensively examined sample trios from 16 neuroblastoma patients: a blood sample as normal control and biopsies from the tumor at initial diagnosis and at relapse, respectively. Among the technologies used were whole-exome sequencing (Illumina HiSeq), array CGH (Agilent CGH arrays), methylation arrays (Illumina HumanMethylation450 BeadChip) and gene expression microarrays (Agilent Custom 44k arrays). While the analysis provided novel insights into the development and evolution of neuroblastoma [SKA+15], it also presented considerable bioinformatics challenges because signals from heterogeneous types of data had to be unified and interpreted.

To identify single nucleotide variants (SNVs) and small indels in each sample, we used our in-house platform Exomate that largely follows the GATK best practice recommendations [VdACH+13] for variant calling and adds an in-house database and interactive web interface for filtering and for variant visualization.

An appropriately filtered variant list from a patient with five relapse samples was the basis for a simple but robust method (Hamming distance between feature vectors combined with neighbor joining) to construct a phylogenetic tree of the initial tumor and the relapses. Even when several filter thresholds were varied considerably, the tree topology stayed constant with bootstrap support values of 100%.

Variants exported from Exomate with coverage and support information formed the basis of an analysis to detect changes in variant allele frequencies between initial tumor and relapses. A two-dimensional scatterplot comparing these frequencies revealed both expected and surprising patterns.

Copy number changes were inferred using Agilent CNV 105k and 180k arrays, displaying higher numbers of genomic aberrations at relapse. While promoter methylation patterns were consistent and patient specific, expression profiles showed clear differences and indicate an involvement of Hippo-YAP signaling.

On the poster we showcase the corresponding data analysis processes and highlight some of the results that were recently discussed in detail [SKA+15].

References

[SKA+15] A. Schramm, J. Köster, Y. Assenov, K. Althoff, M. Peifer, E. Mahlow, A. Odersky, D. Beisser, C. Ernst, A. G. Henssen, H. Stephan, C. Schroder, L. Heukamp, A. Engesser, Y. Kahlert, J. Theissen, B. Hero, F. Roels, J. Altmuller, P. Nurnberg, K. Astrahantseff, C. Gloeckner, K. De Preter, C. Plass, S. Lee, H. N. Lode, K. O. Henrich, M. Gartlgruber, F. Speleman, P. Schmezer, F. Westermann, S. Rahmann, M. Fischer, A. Eggert, and J. H. Schulte. Mutational dynamics between primary and relapse neuroblastomas. *Nature Genetics*, Jun 2015. Advance online access.

[VdACH+13] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*, pages 11.10.1–11.10.33. John Wiley & Sons, Inc., 2013.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Rapid Phylogeny Reconstruction with Support Values

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

Phylogeny construction is usually based on multiple sequence alignments (MSA), which are difficult to compute for genome-scale data. Methods that do not rely on a MSA can be used to compute pairwise distances much more efficiently than the fastest alignment methods available. For example, our recently published program *andi* can compute trees for thousands of bacterial genomes in just hours on desktop-grade hardware [HKP15]. The resulting phylogenies are highly accurate when compared with traditional alignment-based methods. However, these comparisons cannot be done when no reference tree exists. Moreover, without an MSA, bootstrapping cannot be applied to compute support values [Fel85]. We therefore explore the usefulness of other methods for assessing the reliability of individual clades, such as *pairwise* bootstrapping or the quality measurement proposed by Fitch-Margoliash [FM67]. Establishing a good quality measure for trees computed from pairwise distances would have wide applicability among alignment-free phylogeny reconstruction methods.

References

[Fel85] Joseph Felsenstein. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791, 1985.

[FM67] Walter M. Fitch and Emanuel Margoliash. Construction of Phylogenetic Trees. *Science*, 155(3760):279–284, 1967.

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Prediction of Heterosis in Brassica napus by Modeling with a Gene Regulatory Network

Christian Rockmann, Philipp Franke, Heike Pospisil
University of Applied Science Wildau
pospisil@th-wildau.de

Rapeseed (*Brassica napus*) is the most important oilseed crop in Europe and the second most important source of vegetable oil worldwide after soybean [HSB+06]. Spontaneous hybridisation between turnip rape (*B. rapa*) and cabbage (*B.oleracae*) and chromosome doubling formed an allopolyploid genome [CDL+14].

The hybridisation of two different homozygous genotypes can cause F1 hybrids with superior performance relative to their parants. This effect is called heterosis and is successful applied in plant breeding of many crops. However, the genetic mechanism of heterosis in rapeseed and its manifestation under changing enviroment conditions is not sufficiently well known [RBE08]. Heretofore, the parents are selected by empirical breeding. This is an expensive, time consuming process.

The e:Bio project PROGReSs aims at examination of heterosis in Rapeseed using a systems biology approach. Ten universities, research institutes and breeding companies are involved in this interdisciplinary project. Large quantities of genomic, phenotypic and enviromental data are iteratively combined to develop a prediction model for heterosis and yield. Breeders could use this as selection tool to predict performance and exhaust the potential yield of rapeseed hybrids under different conditions.

We will create a gene regulatory network (GRN), which will be integrated in the heterosis model beside phenotypic and enviromental models. With the help of a reference transcriptome and a heterosis simulation model [ERP15] representative amplicons will be selected, which will be sequenced for 662 genotypes. For the de novo transcriptome sequencing we established an RNA-Seq method to get 400bp long reads from the Ion Torrent platform, that enables the correct transcriptome assembly for the two closely related progenitor subgenomes in *Brassica napus* [HKJ+08, CTD+09]. All gene expression data will be integrated with epigenetic profiles from bisulfite sequencing and sRNA sequencing data into a GRN based on statistical models and newly developed algorithms.

References

[CDL+14] B. Chalhoub, F. Denoeud, S. Liu, I.A.P. Parkin, H. Tang, X. Wang, J. Chiquet, H. Belcram, C. Tong, B. Samans, M. Correa, C. Da Silva, J. Just, C. Falentin, C. S. Koh, et al. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, 345(6199):950–953, 2014.

[CTD+09] F. Cheung, M. Trick, N. Drou, Y.P. Lim, J.-Y. Park, S.-J. Kwon, J.-A Kim, R. Scott, J.C. Pires, A.H. Paterson, C. Town, and I. Bancroft. Comparative analysis between homoeologous genome segments of Brassica napus and its progenitor species reveals extensive sequence-level divergence. *The Plant cell*, 21(7):1912–1928, 2009.

[ERP15] P. Emmrich, H. Roberts, and V. Pancaldi. A Boolean gene regulatory model of heterosis and speciation. *BMC Evolutionary Biology*, 15(1):24, 2015.

[HKJ+08] E.C. Howell, M.J. Kearsey, G.H. Jones, G.J. King, and S.J. Armstrong. A and C genome distinction and chromosome identification in Brassica napus by sequential fluorescence in situ hybridization and genomic in situ hybridization. *Genetics*, 180(4):1849–1857, 2008.

[HSB+06] M. Hasan, F. Seyis, a. G. Badani, J. Pons-Kühnemann, W. Friedt, W. Lühs, and R. J. Snowdon. Analysis of genetic diversity in the Brassica napus L. gene pool using SSR markers. *Genetic Resources and Crop Evolution*, 53(4):793–802, 2006.

[RBE08] M. Radoev, H.C. Becker, and W. Ecke. Genetic analysis of heterosis for yield and yield components in rapeseed (Brassica napus L.) by quantitative trait locus mapping. *Genetics*, 179(3):1547–1558, 2008.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Cancer Specific Copy Number Variations

Stefanie Marczok, Birgit Bortz, Heike Pospisil
University of Applied Science Wildau
pospisil@th-wildau.de

The number of cancer cases worldwide is still increasing and reached 14 million (new events) in 2012 [VPV+13] [SW14]. This fact can not only be explained by the obvious reason (our increasing lifespan), but is also attributable to the poor knowledge about tumorigenesis, tumor progression and insufficient therapies. In the last few years, the focus of cancer research moved to structural genomic variations caused by ineffective and incorrect repair mechanisms [GFJ08, HLRI9b]. Copy number variations (CNVs) as some kind of structural variations are charaterised by deletions or amplifications of genomic regions [FCS06] and may lead to disfunctioned genes or to an altered gene expression [PA05].

First investigations of our group revealed that the appearance and the frequency of CNVs are highly correlated to the progressive stage of cancer [SPK12]. To determine the influence of CNVs to cancer, we identified, compared and characterized CNVs in a public accessible SNP array (Affymetrix Genome-Wide Human SNP Array 6.0) data set of 2,820 different cancer genomes from eight tumor entities and from 432 healthy genomes for comparison reasons. We identified CNVs comprising known tumor suppressor and oncogenes. Further, we were especially interested in the boundaries between two regions of significantly different copy numbers (so called breakpoint regions) and found 31 of them which seem to be cancer specific. They were further be classiffed in tumor entity :specific, cancer specific and common breakpoint regions. As far as we know it is the first comprehensive and comparative CNV analysis for several tumor entities and it can serve as a starting point for further clarification of the tumorigenesis and progression [WCM+13, WJS+14].

References

[FCS06] L. Feuk, A.R. Carson, and Scherer S.W. Structural variation in the human genome. *Nature*, 7(2):85–97, Feb 2006.

[GFJ08] W. Gu, Zhang F., and Lupski J.R. Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(4), 2008.

[HLRI9b] P.J. Hastings, J.R. Lupski, S.M. Rosenberg, and G. Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10:551–564, May 2009b.

[PA05] D. Pinkel and D.G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37:S11–S14, May 2005.

[SPK12] C. Standfuss, H. Pospisil, and A. Klein. SNP microarray analyses reveal copy number alterations and progressive genome reorganization during tumor development in SVT/t driven mice breast cancer. *BMC Cancer*, 12:380, 2012.

[SW14] B.W. Stewart and C.P. Wild. In *World Cancer Report 2014*, chapter Cancer etiology. World Health Organization, 2014.

[VPV+13] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, Mar 2013.

[WCM+13] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J.M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.*, 45(10):1113–1120, Oct 2013.

[WJS+14] N. Weinhold, A. Jacobsen, N. Schultz, C. Sander, and W. Lee. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet.*, 46(11):1160–1165, Nov 2014.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Prioritizing cancer driver genes with ContrastRank

Rui Tian, Malay K. Basu, Emidio Capriotti*
*Institute for Mathematical Modeling of Biological Systems, University of Düsseldorf
emidio.capriotti@hhu.de

One of the main challenges in personalized medicine is the interpretation of the sequence variations in the human genome [FG11]. The recent advance in high-throughput sequencing is generating a huge amount of data, which are important resources for deciphering the genotypes underlying a given phenotype. Recently genome sequencing has been extensively applied to cancer genome studies as well. The Cancer Genome Atlas (TCGA) Consortium is making available genomic data for >30 cancer types (http://cancergenome.broadinstitute.org/) to the scientific community, enabling the development of computational approaches to define a profile of the genetic variations associated with specific cancer samples. These variation profiles can be used to prioritize cancer-associated genes and classify cancer genotypes [TR15].

In this work we present ContastRank, a new method for the prioritization of putative impaired genes in cancer. The method is based on the comparison of the putative defective rate of each gene in tumor versus that in normal and 1000 Genomes samples [Con10]. We show that the method is able to provide a ranked list of putative impaired genes for colon, lung and prostate. The list significantly overlaps with the list of known cancer driver genes previously published (http://cancergenome.broadinstitute.org/). More importantly, by using our scoring approach, we can successfully discriminate between TCGA normal and tumor samples in our datasets.

Our binary classifier based on ContrastRank score reaches an overall accuracy higher than 90% and the Area Under the Curve (AUC) of Receiver Operating Characteristics (ROC) higher than 0.95 for all the three types of adenocarcinoma analyzed in this paper. In addition, using ContrastRank score we are able to discriminate the three tumor types with a minimum overall accuracy of 77% and AUC of 0.83 [TR14].

A web server implementation of the method is available at:
http://snps.biofold.org/contrastrank

References

[Con10] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 473:1061–1073, 2010.

[FG11] Daneshjou R Karczewski KJ Altman RB Fernald GH, Capriotti E. Bioinformatics challenges for personalized medicine. Bioinformatics. volume 37, pages 1741–1748. 2011.

[TR14] Capriotti E. Tian R, Basu MK. ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples. *Bioinformatics*, 30:i572–i578, 2014.

[TR15] Capriotti E. Tian R, Basu MK. Computational methods and resources for the interpretation of genomic variants in cancer. *BMC Genomics*, 16 (Suppl. 8):S7, 2015.

Poster 36

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Recent development of the phylogenetic tree editor TreeGraph 2

Ben C. Stöver, Sarah Wiechers and Kai F. Müller
Evolution and Biodiversity of Plants Group, Institute for Evolution and Biodiversity, WWU Münster
stoever@bioinfweb.info

TreeGraph 2 is a user friendly and widely used tree editor with the main focus on processing, visualizing and comparing phylogenetic trees carrying numerous annotations. Since its initial publication [SM10] a number of features have been added.

The ability to import and visualize probabilities for ancestral character states (as reconstructed by software packages like *BayesTraits* [MPB04]) e.g. by pie charts attached to internal nodes has been added. A special reader allows importing according data from the *BayesTraits*-specific format, while importing from similar software is indirectly possible using the new and more powerful annotation table import function.

In addition to the published feature that compares conflicting support values from alternative trees, a new visual comparison method allows the user to directly investigate topological and support differences. If nodes in one tree are selected, TreeGraph 2 directly visualizes according nodes or regions with conflicting topologies and support in other opened trees.

Calculating branch and node annotations from each other using custom expressions has been extended by several new functions which e.g. allow whole columns (one type of annotation from all nodes in a tree) or rows (all annotations on one node) as input. Additional new features include the closest possible sorting of terminal nodes according to a specified order, automatic collapsing of internal nodes depending on annotations (e.g. support) or rerooting trees by a given outgroup which may be in topological conflict with the tree.

The poster introduces the new features of TreeGraph 2 and shows examples for their application in a recent project investigating the influence of different automated and manual multiple sequence alignments of non-coding sequences containing certain microstructural mutations on phylogenetic inference.

Software and poster download and documentation: http://treegraph.bioinfweb.info/

References

[MPB04] Andrew Meade Mark Pagel and Daniel Barker. Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst Biol*, 53(5):673–684, 2004.

[SM10] Ben C. Stöver and Kai F. Müller. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*, 11(7), 2010.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

How good is Flux Balance Analysis? A comparison with nonlinear simulations of a simplified whole-cell model

Hugo Dourado and Martin J. Lercher
Institute for Computer Science, Heinrich Heine University Düsseldorf
hugo.dourado@hhu.de

Based on reaction stoichiometries and linear optimization, Flux Balance Analysis (FBA) is widely used to simulate genome-scale metabolism. However, FBA fails to predict some important metabolic phenomena related to physical constraints other than stoichiometry, e.g., the shift from high yield to low yield pathways when nutrients are abundant (overflow metabolism) [NT76].

To go beyond FBA, other important cellular constraints must be considered, such as the limitation in volume for cell components (e.g. ribosomes, enzymes and metabolites), the limitation in surface area for transporters, the costs of protein production, and the reaction kinetics dependent on enzyme and metabolite concentrations. Fully accounting for those constraints requires computationally inefficient nonlinear optimizations as well as complete knowledge of kinetic parameters, potentially forbidding the simulation of genome-scale models.

In order to assess the importance of these constraints in predicting cellular growth and reaction rates, we compared nonlinear simulations of a simplified whole-cell metabolic model (based on previous work [MvBdRT09]) to FBA and to FBA with molecular crowding. FBA growth rate predictions are realistic only at low growth rates, while FBA with molecular crowding provides qualitatively acceptable predictions at all growth rates. However, the flux distributions predicted by both models differ significantly from the nonlinear simulations.

References

[MvBdRT09] D. Molenaar, R. van Berlo, D. de Ridder, and B. Teusink. Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular Systems Biology*, 5(1):323, 2009.

[NT76] O. M. Neijssel and D.W. Tempest. The role of energy-spilling reactions in the growth of Klebsiella aerogenes. *Archives of Microbiology*, 110(1):305–311, 1976.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Evaluation of LOH detection approaches on whole-genome and exome sequencing data of 30 retinoblastoma patients

Corinna Ernst[1], Christopher Schröder[1], Petra Temming[2,3,4] and Sven Rahmann[1]

[1]Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, Germany
[2]Clinic for Paediatrics III, University Hospital Essen, Germany
[3]Eye Cancer Research Group, Institute of Human Genetics, University of Duisburg-Essen, Germany
[4]German Consortium for Translational Cancer Research (DKTK), Heidelberg, Germany,

corinna.ernst@uni-due.de

Loss of heterozygosity (LOH) serves as a general term for the transformation of a heterozygote allele into a homozygote – respectively hemizygote – one, and is associated with the development of various tumors, e.g. retinoblastoma, ovarian carcinoma or acute myeloid leukemia [RDG+15]. However, although widely used in the literature, the term lacks accuracy as it does not specify a variant's genotypic state and the underlying mechanism by which it was created. Various tools claiming reliable identification of regions of LOH using exome and/or whole-genome sequencing data were published in recent years. We compare and validate these approaches based on an outstanding data set of whole-genome and exome sequencing samples of 30 retinoblastoma patients. In particular, we focus on the identification of regions of LOH overlapping gene RB1 on chromosome 13. Inactivation of both alleles of RB1 is a well-studied mechanism in retinoblastoma development [HD99, LGB+97].

References

[HD99] SA Hagstrom and TP Dryja. Mitotic recombination map of 13cen-13q14 derived from an investigation of loss of heterozygosity in retinoblastomas. *Proc Natl Acad Sci U S A*, 96(6):2952–7, 1999.

[LGB+97] DR Lohmann, M Gerick, B Brandt, U Oelschlger, B Lorenz, E Passarge, and B Horsthemke. Constitutional RB1-gene mutations in patients with isolated unilateral retinoblastoma. *Am J Hum Genet*, 61(2):282–294, 1997.

[RDG+15] GL Ryland, MA Doyle, D Goode, SE Boyle, DY Choong, SM Rowley, J Li, Australian Ovarian Cancer Study Group, DD Bowtell, RW Tothill, IG Campbell, and KL Gorringe. Loss of heterozygosity: what is it good for? *BMC Medical Genomics*, 8(1):45, 2015.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Markov-Chain Monte-Carlo Sampling of Metabolite Concentrations to Identify Thermodynamically Feasible Reaction Directionalities for Flux Balance Analysis

Ulrich Wittelsbürger, Katrin Schrankel and Martin J. Lercher
Bioinformatics Institute, Heinrich Heine University Düsseldorf
ulrich.wittelsbuerger@hhu.de

Flux balance analysis (FBA) is a widely used tool for both the understanding and design of cell metabolism. Without any kinetic data, qualitative as well as quantitative predictions of metabolic flux activity to achieve optimal growth can be made. It is also computationally inexpensive as only linear problems need to be solved. Among its shortcomings, however, is the lack of a rigorous enforcement of the thermodynamical feasibility of reactions; this is due to the fact that FBAs steady state assumption evades the consideration of metabolite concentrations, which impact changes in Gibbs' free energies.

Existing approaches to ensure thermodynamic feasibility are based on computationally expensive mixed integer linear programming. Here, we propose a Markov Chain Monte Carlo (MCMC) approach to identify metabolite concentrations that ensure thermodynamic feasibility.

We sample metabolite concentration vectors and use these to determine reaction directionalities that violate thermodynamic constraints. Using FBA we then identify a biomass producing solution with minimal total flux F through these infeasible reactions. Solutions with lower F are deemed more likely, and hence the sampled concentrations are expected to converge towards thermodynamically feasible solutions.

We show propose and evaluate several MCMC designs on metabolic networks of different sizes.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Metabolic modelling reveals evolution of complex phenotypes through successions of adaptive steps

Tin Y Pang and Martin Lercher
Institute for Bioinformatics, Heinrich Heine University Duesseldorf
lercher@cs.uni-duesseldorf.de

Comparing the metabolic networks of closely related bacterial strains can reveal their adaptive history. We analyzed 53 *E. coli* strains with published metabolic network models [MCA+13] and reconstructed their phylogeny. We considered each of the 52 internal nodes of the tree as an ancestral strain, and reconstructed the ancestral genomes and metabolic networks using maximum parsimony. We then identified the enzymes horizontally transferred between the ancestral strains and grouped them into segments according to their physical proximity in the 53 genome sequences.

We searched for the benefits brought by the transfer of those segments to the host metabolic networks, *i.e.*, the new phenotypes acquired and the existing phenotypes enhanced, using flux balance analysis (implemented in the sybil R package [GDDFL13]). We compared the benefits of those segments with a model that simulates the transfer of all possible genomic segments between different *E. coli* strains, as well as with another model that simulates transfer of randomized segments. Our analysis shows that, compared with randomized segments, real genomic segments are more likely to transfer reactions and phenotypes.

We further examined how a bacterium can acquire a complex phenotype, *i.e.*, a phenotype whose genes are distributed across multiple locations on the genome, requiring transfer of more than a single segment. For each transferred or enhanced phenotype in a given strain, we searched for the contributing segments added since the strains most recent ancestor, and found that almost all phenotypic gains can be explained by the transfer of a single segment. However, if we trace back farther in time and identify the segments contributing to a given phenotype that were added along the phylogenetic lineage (i.e., after the most recent common ancestor of the 53 strains), then we found that a substantial fraction of phenotypic benefits (40%) required multiple transfers. Thus, the most recent segment transfer required the presence of previous transfers to be beneficial. This is evidence of preadaptation (or exaptation): in order to adapt to an environment that requires complex metabolic changes, a bacterium may have to take an evolutionary detour through intermediate environments to acquire segments one-by-one adaptively.

References

[GDDFL13] Gabriel Gelius-Dietrich, Abdelmoneim A. Desouki, Claus J. Fritzemeier, and Martin J. Lercher. sybil â Efficient constraint-based modelling in R. *BMC Systems Biology*, 7(1):125–8, November 2013.

[MCA+13] Jonathan M. Monk, Pep Charusanti, Ramy K. Aziz, Joshua A. Lerman, Ned Premyodhin, Jeffrey D. Orth, Adam M. Feist, and Bernhard Ø. Palsson. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences of the United States of America*, 110(50):20338–20343, December 2013.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Unifying open chromatin assays using supervised learning for transcription factor prediction

Florian Schmidt[1,3,4], Jonas Fischer[1,3], Karl J Nordstroem[2], Nina Gasparoni[2], Gilles Gasparoni[2], Kathrin Kattler[2], Nico Pfeifer[1], Joern Walter[2], Marcel H Schulz[1,3]
[1] *Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, Germany* [2] *Department of Genetics, Saarland University, Saarbrücken, Germany* [3] *Cluster of Excellence on Multimodal Computing and Interaction, Saarland University, Saarbrücken, Germany and* [4] *Graduate School of Computer Science, Saarland University, Saarbrücken, Germany*
fschmidt@mmci.uni-saarland.de

Transcription factors (TFs) play an essential role in gene regulation and are involved in many different diseases. One common approach for predicting transcription factor binding is to measure genome-wide open chromatin regions and then use an integrative analysis method to combine TF motif information and other auxiliary data, e.g. histone ChIP-seq.

Different assays have been proposed to measure open chromatin genome-wide. We compare three of these techniques, NOMe, DNase1 and ATAC-seq to understand differences and commonalities between them. Despite a significant overlap, each method has its own advantages in finding open chromatin regions resulting in method specific assessment of open chromatin. Thus, predicting transcription factor binding in open chromatin regions defined by different assays can lead to contradictory results. This discrepancy could be reduced by improving open chromatin assessment, either through optimization of assay protocols or by computational post processing of the experimental results.

Here, we propose an efficient computational pipeline that combines an SVM-based open chromatin classifier with motif based TF affinity predictions [RKMV07]. We generate a comprehensive gold standard data set of open chromatin regions using DNase1, NOME, and ATAC-seq data generated in the DEEP project, that we use for testing and optimising our pipeline. For validation, we use ChIP-seq TF data from ENCODE.

Applying the general pipeline on datasets from different open chromatin assays can help to remove disagreement between the aforementioned methods.

References

[RKMV07] H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan 2007.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

dinopy: Efficient DNA Input and Output with Cython

Henning Timm and Till Hartmann
Genome Informatics, Institute of Human Genetics, Faculty of Medicine, University Hospital Essen, University of Duisburg-Essen, 45122 Essen
henning.timm@tu-dortmund.de

Dinopy is a Python package that aims to simplify the development of bioinformatics applications by providing efficient facilities for DNA input and output.

When developing a Python application that works on DNA sequences, file IO sooner or later becomes an issue. A parser for FASTQ files for example, is easily written, but this is an error prone task that can slow down both the development process and the runtime of the application. Using one of the established libraries works fine, until a different data type is required. This forces the developer to manually convert the input data, which can needlessly slow down the entire application. Another problem is, that established libraries like Biopython [CAC$^+$09] often favor generality over performance. Dinopy (Dna INput and Output in PYthon) aims to eliminate this time consuming factors by providing input and output facilities specialized on the efficient interaction with FASTA and FASTQ files.

To achieve this, dinopy uses a data type system comparable to the one used by numpy [VDWCV11]. In addition to this dinopy is specialized on DNA sequences, what allows us to use highly optimized and tailored facilities. For an additional gain of speed, dinopy has been implemented in Cython [BBC$^+$11] and uses extension modules implemented in C and C++.

On the analysis side dinopy offers several levels of access to the given data. For example a genome can be accessed chromosome by chromosome or by all of its q-grams, depending on the needs of the user. To keep a low memory profile while working with a steadily increasing amount of data, most of dinopy's facilities are implemented as generators.

In addition to its capabilities as a file parser and writer dinopy also provides several processors for rapid prototyping as well as productive use. These include (gapped) q-grams and a suffix array using a linear time construction algorithm [NZC11].

References

[BBC$^+$11] Stefan Behnel, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith. Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31–39, 2011.

[CAC$^+$09] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[NZC11] Ge Nong, Sen Zhang, and Wai Hong Chan. Two efficient algorithms for linear time suffix array construction. *Computers, IEEE Transactions on*, 60(10):1471–1484, 2011.

[VDWCV11] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

CompGSA: A Comparability Approach using Gene Set Analysis and Trimmed Linear Regression

Yang Xiang and Florian Martin
Department of Biological System Research, Philip Morris International R&D, Neuchatel, Switzerland
yang.xiang@pmi.com and florian.martin@pmi.com

Many methods of gene set analysis, e.g., GSEA [STM$^+$05], GSA [ET07], and Globaltest [GVDGDKVH04], have been developed in recent years. Additionally to extract relevant biological processes from the data in hand, they have been widely used for exploring comparability of different data sets or contrasts (e.g., treatment vs. control effects). For example, Beane *et al* [BSL$^+$07] used GSEA to check the similarity between their data set and other public data sets, by testing the enrichment of their differentially expressed gene list in the other datasets and by comparing the genesets results between contrasts.

The latter aspect is motivating our work, where a stronger comparability approach is defined. Additionally, comparing contrasts within a study is mandatory in systems toxicology experiments whereby doses and exposure duration are systematically compared.

Significantly up- or down-regulated gene set (self-contained null hypothesis) in two contrasts does not guarantee that these contrasts are comparable on this gene set as the associated gene fold-changes may behave very differently. Furthermore, even if a geneset is not significant, a subset of substantial size may lead to significance and high similarity between the corresponding gene fold-changes. This is a typical behavior for geneset describing pathways in which several alternative signaling paths are present.

For any predefined gene set A and two contrasts C_1 and C_2, a novel method, *CompGSA*, which aim at identifying a maximal subset $A^* \subseteq A$ for which C_1 and C_2 have similar observed fold-changes for the genes in A^* and that is significantly perturbed in both C_1 and C_2, was developed. Trimmed linear regression was used as the key tool for solving this problem. The application of *CompGSA* in simulated data and experimental data demonstrated the added value provided by our approach.

References

[BSL$^+$07] Jennifer Beane, Paola Sebastiani, Gang Liu, Jerome S Brody, Marc E Lenburg, and Avrum Spira. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol*, 8(9):R201, 2007.

[ET07] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The annals of applied statistics*, pages 107–129, 2007.

[GVDGDKVH04] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[STM$^+$05] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Image processing algorithms for online monitoring of cell density in Pichia pastoris cultivations by In situ Microscopy.

D. Marquard, A. Enders, G. Roth, P. Lindner, T. Scheper
Institut fuer Technische Chemie, Leibniz University Hannover
marquard@iftc.uni-hannover.de

In situ Microscopy (ISM) is an optical non-invasive method to monitor processes in real-time. *Pichia pastoris* is one of the most promising protein expression systems. This yeast combines prokaryotic and eukaryotic features. *Pichia pastoris* can reach high cell densities, which is a big challenge for online monitoring systems and especially for the image processing algorithms.

In this poster two algorithms for determining cell density from ISM images are displayed. The first algorithm detects single and clustered *Pichia pastoris* cells and counts them. This algorithm has good accuracy at lower cell densities, but at higher cell density the algorithm loses accuracy. Especially if the cluster formation is very strong the algorithm fails completely. The second algorithm determines the size of clusters formed by the *Pichia pastoris* cells in the images. With this algorithm the cell densities of almost the whole cultivation can be monitored, as long as clusters are formed by the cells. Right at the beginning of all cultivations and when the cells grow very slowly for any reason, almost no clusters can be recognized. A combination of both algorithms enables a full range cell density online monitoring for *Pichia pastoris* cultivations up to 75 g/L bio dry mass or an optical density of 200 without seeing any effects of saturation in the correlations.

Overall this poster shows that ISM in combination with suitable image processing algorithm is a strong tool for real-time monitoring of cell density in *Pichia pastoris* cultivations.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

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

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

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

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

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

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Mining for common reactivity patterns of human autoantibodies against endogenous protein targets using clustered autoantibody reactivities

HD. Zucht[1], D. Chamrad[1], A. Telaar[1], H. Göhler[1], M. Gamer[1], S. Vordenbäumen[2], P. Schulz-Knappe[1], M. Schneider[2], P. Budde[1]
[1]*Protagen AG, Dortmund, Germany,* [2]*Centre of Rheumatology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany*
hans-dieter.zucht@protagen.com

Abstract

Autoimmune diseases arise from an abnormal immune response of the body against self-proteins leading to tissue and organ damage. The excessive production of harmful autoantibodies (AAB) is a hallmark of autoimmune diseases including Rheumatoid Arthritis (RA), Systemic Lupus Erythematodes (SLE), Systemic Sclerosis (SSc) and Sjögren's Disease (SjS). Early diagnosis of autoimmune diseases is essential for initiation of therapy, but often challenging due to the heterogeneous clinical presentation of the diseases. Aim: AABs serve as diagnostic markers for various autoimmune diseases, but the co-occurrence of AABs in SLE patients has rarely been analyzed. Detecting a broad set of AABs might help to investigate the number, co-prevalence and similarities of AAB reactivities in autoimmune diseases such as SLE. Here, we describe the cluster analysis of a multiplexed AAB array of 87 AABs in a variety of disease entities.

Methods

A Luminex bead-based AAB assay has been developed comprising established and novel biomarkers for differential diagnosis of SLE, markers for disease activity, organ involvement, interferon type I dependent reactivities and novel biomarkers. AAB reactivity against these antigens was tested in over 700 SLE, healthy controls (n=1000), and other autoimmune diseased patient samples (n=500). Data analysis was based on biclustering algorithms, Kohonen mapping of antibody reactivity. Cluster analysis was also performed using transformed datasest (qualitative) to investigate and visualize characteristic marker prevalence and co-prevalence patterns.

Results

Based on the individual marker pattern, patients either belong to clusters defined by characteristic markers, or are phenotypically more overlapping with each other. Systemic sclerosis patients clearly split up into two patient clusters, whith different AAB reactivity. SLE patients can be mapped to at least four different reactivity groups (G1-G4) including patients: G1) a higher disease activity score, broad and homogeneous AAB reactivity; G2) with broad, but heterogeneous AAB reactivity; G3) who have few AABs and G4) with unusual AAB patterns.

Conclusion

The multiplexed analysis of AABs in autoimmune diseases enables to investigate disease subgroups. The utility of this approach is to support the stratification of patients for an improved therapy in a personalized medicine approach.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Prediction of Specific Transcription Factor Binding Sites from DNase-seq data using Hidden Markov Model

Michael Rauer[1,*], Mathieu Clément-Ziza[1], Achim Tresch[2], Uwe Ohler[3]
[1] *ZMMK,Universität zu Köln, 50931 Köln*
[2] *Biozentrum, Universität zu Köln, 50674 Köln*
[3] *Max-Delbrück-Centrum für Molekulare Medizin, 13125 Berlin*
[*] michael.rauer@uni-koeln.de

In regulatory genomics, the identification of the functional binding sites of several transcription factors (TFs) can be pivotal in understanding the regulation of gene expression in different conditions. However, performing this task in multiple conditions is experimentally complicated and demanding. This problem can be addressed by using condition specific DNase I hypersensitive sites sequencing (DNase-seq). DNase-seq is a method to identify regions of open chromatin [SC10]. Furthermore, DNase I footprints for specific TFs can be found at high resolution. Here we propose an algorithm for the joint modeling of DNase I signal and DNA sequence, to predict condition-specific binding sites of multiple TFs at single nucleotide resolution.

At low resolution (ca. 500 nt) clusters of DNase-seq reads indicate regions of open chromatin (DNaseI hypersensitivity sites). Investigating DNase-seq signal at single nucleotide resolution reveals footprints (ca. 5 - 25 nt), which indicate TF binding [BSL+11]. It has been shown that the profile of these footprints can be TF specific [YFCO14].

Here, we present a TF-specific model that predicts TF binding jointly from nucleotide-sequence and DNase footprints. Using only DNase-seq read counts and DNA sequence, we applied a bidirectional multivariate Hidden Markov Model [ZLC+14] to model TF binding. The model locates potential TF binding sites (one model per TF) and identifies if they are bound based on the underlying DNase profile. We tested the model for the following TFs: CTCF, REST, STAT1, YY1, as well as MYC and MAX. We used the corresponding ChIP-seq experiments, to evaluate the accuracy of our predictions of bound TF binding sites. Our results indicate, that the prediction of bound TF binding sites improves over an nucleotide sequence-based approach (MEME-FIMO [GBN11]). Furthermore, we demonstrate that we can reuse our previously trained models to predict bound TF binding sites in other DNase-seq experiments. Thus, we can predict TF binding sites bound under certain conditions, without performing further ChIP-seq experiments.

References

[BSL+11] AP Boyle, L Song, B Lee, D London, D Keefe, E Birney, VR Iyer, GE Crawford, and TS Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, 21(3):456–64, March 2011.

[GBN11] CE Grant, TL Bailey, and WS Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, 27(7):1017–8, April 2011.

[SC10] L Song and GE Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, 2010(2):pdb.prot5384, February 2010.

[YFCO14] GG Yardimci, CL Frank, GE Crawford, and U Ohler. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.*, 42(19):11865–78, October 2014.

[ZLC+14] B Zacher, M Lidschreiber, P Cramer, J Gagneur, and A Tresch. Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. *Mol. Syst. Biol.*, 10(12):768, January 2014.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Inferring Directional Genetic Interactions from Mutli-Parametric Combinatorial Perturbation Screens

Bernd Fischer
Computational Genome Biology, German Cancer Research Center
b.fischer@dkfz.de

Genes display epistatic (genetic) interactions, whereby the presence of one genetic variant can mask, alleviate or amplify the phenotypic effect of other variants. Such a directional relationship is present, for instance, if one gene product positively or negatively regulates the activity of the other, if its function temporally precedes that of the other, or if its function is a necessary requirement for the action of the other. Large-scale synthetic genetic interaction screens have been performed and have been predictive for functional relationships between genes in yeast, E. coli, C. elegans and metazoan cells. To date, all large-scale genetic interaction studies have been designed to detect gene-gene interactions based on the definition of an interaction as a departure from the combination of the genes individual effects. This statistical definition provides limited information on the directional relationship between the genes. We will present a method that combines genetic interactions on multiple phenotypes to reveal directional relationships, and report a dense regulatory network covering 1367 genes [FSH+15]. It reveals the directional, temporal and logical relationships between genes and allows us to dissect regulatory networks using high-throughput intervention experimentation. The network could reconstruct the sequence of protein activities in mitosis, and revealed that the Ras pathway interacts with the SWI/SNF chromatin-remodelling complex, which we show is conserved in human cancer cells. Our work presents a powerful approach for reconstructing directional regulatory networks, and provides a resource for the interpretation of functional consequences of genomic alterations in disease.

[FSH+15] [HSF+11] [LFB+13]

References

[FSH+15] Bernd Fischer, Thomas Sandmann, Thomas Horn, Maximilian Billmann, Varun Chaudhary, Wolfgang Huber, and Michael Boutros. A map of directional genetic interactions in a metazoan cell. *eLife*, 4:e05464, 2015.

[HSF+11] Thomas Horn, Thomas Sandmann, Bernd Fischer, Elin Axelsson, Wolfgang Huber, and Michael Boutros. Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nature methods*, 8(4):341–346, 2011.

[LFB+13] Christina Laufer, Bernd Fischer, Maximilian Billmann, Wolfgang Huber, and Michael Boutros. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nature methods*, 10(5):427–431, 2013.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Dynamic updates of reference improve NGS read mapping

Karel Břinda, Valentina Boeva, and Gregory Kucherov
LIGM Université Paris-Est and Institut Curie
karel.brinda@univ-mlv.fr

We present results demonstrating that *dynamic read mapping*, i.e., mapping with updating reference in accordance to already mapped reads, provides better accuracy and sensitivity than its static analog. We also describe scenarios where the improvement over *static mappers* is the most significant.

In spite of a great attention payed to NGS read mapping, *dynamic mapping* has not been well-studied yet. Existing literature includes only a semestral project of Jacob Pritt [Pri13] containing several interesting observations and ideas, and a program called Dynmap [IKM+12].

Firstly, dynamic mappers bring several technical and algorithmic issues, above all: *i)* underlying data structures (FM indexes, hash-tables, etc.) must be dynamic; *ii)* expensive statistics about already mapped reads must be kept in memory during whole mapping, *iii)* addressing in the reference must be somehow generalized, and *iv)* remapping and unmapping already mapped reads should be considered. Inevitably, these four points imply that the memory consumption will extensively increase while the performance will decrease, in comparison to *static mappers*. Secondly, *dynamic mapping* could be beneficial to a limited class of applications when speed can be traded for sensitivity and selectivity (far reference, many mutation hotspot regions, etc.).

Besides the above mentioned approaches (*static mapping* and *dynamic mapping*), there exists an intermediate approach, a so-called *iterative referencing* [GG13], which is based on iterative repetition of '*static mapping* of all reads' and 'consensus calling' until number of updates decreases below a given threshold. While *dynamic mappers* update the reference having only partial information at their disposal, *iterative referencing* works with full information about alignments of all reads, therefore, it provides better results.

In order to examine in what situations *dynamic mappers i)* are superior over *static mappers*, and *ii)* can approach results of *iterative referencing*, we developed a pipeline called DyMaS (Dynamic Mapping Simulator), which uses selected state-of-the-art *static mappers* to map reads in small batches followed by computation of statistics and update of the reference sequence. Even though *dynamic mappers* should be capable to perform updates after mapping each new read, we can reach a good approximation of their behavior with our tool when batches of reads are small enough.

In our work, we compare *static mapping*, *iterative referencing*, and *dynamic mapping* under several scenarios (various rates of genomic mutations, sequencing errors, and sample contamination). The obtained improvements are demonstrated on 'precision-sensitivity' diagrams produced using RNFtools [BBK15].

References

[BBK15] K. Břinda, V. Boeva, and G. Kucherov. RNF: a general framework to evaluate NGS read mappers. *arXiv:1504.00556 [q-bio.GN]*, 2015.

[GG13] A. Ghanayim and D. Geiger. Iterative referencing for improving the interpretation of DNA sequence data. Technical report, Technion, Israel, 2013.

[IKM+12] C. S. Iliopoulos, D. Kourie, L. Mouchard, T. K. Musombuka, S. P. Pissis, and C. de Ridder. An algorithm for mapping short reads to a dynamically changing genomic sequence. *Journal of Discrete Algorithms*, 10:15–22, 2012.

[Pri13] J. Pritt. Efficiently Improving the Reference Genome for DNA Read Alignment, 2013.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Ultra-fast phylo-functional classification of metatranscriptomic reads

Heiner Klingenberg and Peter Meinicke
Department of Bioinformatics, Institute of Microbiology and Genetics, University of Göttingen
{heiner,peter}@gobics.de

Metatranscriptomics can yield a direct snapshot of the dynamics of a microbial community. In contrast to metagenomic sequences which just indicate the functional potential of the community, metatranscriptomic RNA-Seq data provides evidence for gene functions and metabolic activities that are actually performed by the community [SD11]. The experimental data consists of a large number of short sequencing reads which have to be classified according to the phylogenetic origin and putative functions. The simultaneous binning into taxonomic and functional categories makes it possible to actually answer the question "who is doing what?" under the given environmental conditions. Classical bioinformatics tools quickly reach computational limits when assigning vast amounts of short reads to taxonomic and functional categories. Our tool UProC [Mei15] offers researchers a very fast protein domain classification, which on short reads (100 bp) outperforms profile-based methods like HMMER and RPS-BLAST. In UProC, the classification of a protein sequence is based on assigning all k-mers (k=18) in the sequence by similarity to functionally labelled words of the same length in a reference database.

For an extension of UProC we have investigated the possibility to add rank-specific taxonomic labels to the database words and use these labels for a phylogenetic classification of sequencing reads. The word-specific taxonomic label can range from species to superkingdom, or no specific category at all, depending on the location of the assigned reference word in the taxonomic tree of the reference protein family. The evaluation of the taxonomic labels that result from all words that could be assigned to the detected protein families is similar to the lowest common ancestor scheme often applied to BLAST results. First results indicate that a reliable phylo-functional classification of metatranscriptomic reads is possible with UProC at a very high speed.

References

[Mei15] P. Meinicke. UProC: tools for ultra-fast protein domain classification. *Bioinformatics*, 31(9):1382–1388, 2015.

[SD11] C. Simon and R. Daniel. Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.*, 77(4):1153–1161, Feb 2011.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Experimentally-based simulation and quantification of transcriptomic high-throughput sequencing

Steffi Kästner, Christian Remmele, Marcus Dittrich, Tobias Müller
Department of Bioinformatics, Biocenter, Univiersity of Würzburg
Tobias.Mueller@biozentrum.uni-wuerzburg.de

RNA-Sequencing is a well-established technique, yet still developing further. The more recently emerging methods, i.e. single molecule sequencing technologies which are able to produce ultra-long reads in the range of 10kbp or more, might pose new challenges in the analysis of RNA-Seq data. To assess this issue on a larger scale, we complemented the Flux-Simulator [GZR+12], an RNA-Seq simulation software able to imitate the sequencing procedure, with an expression profile based on experimentally derived data, and utilized the workflow system snakemake [KR12] to simulate and quantify a large number of data sets. We thereby developed a pipeline that allows for parallelized biology- but also protocol-oriented simulations, and quantication of RNA-Seq experiments, and used this tool to investigate the effects of read length, sequencing depth and single-end sequencing compared to paired-end sequencing.

In order to simulate with an expression profile retaining the characteristics of biological gene expression, we conveyed real biological counts to a transcript expression profile suitable to the Flux-Simulator's format. As RNA-Seq counts are often reported in gene counts, we developed a method to transfer gene counts to transcript counts. The expression distribution for most of the experiments was based on one of the first RNA-Seq data sets, the transcriptome of *Mus musculus* in different tissues from Mortazavi et. al [MWM+08]. We set up an extensive framework relying on snakemake, capable of simultaneously simulating multiple RNA-Seq experiments, and simulated a range of read lengths (mainly relating to Illumina sequencing systems but also included ultra-long reads relevant in particular for third generation sequencing methods) and read numbers, as well as single end and paired end reads. The framework also included parallel quantication of these reads primarily with Tophat2 and Cufflinks from the Tuxedo-Pipeline [KPT+13] [TWP+10]. The results were analyzed in comparison to the original expression and also in terms of differential gene expression.

Our pipeline proved to be a valuable tool to allow insights into the effects of read length and number as well as paired-end sequencing on downstream analysis of RNA-Seq.

References

[GZR+12] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 40(20):10073–83, November 2012.

[KPT+13] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013.

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

[MWM+08] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

[TWP+10] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

An Information Theoretic View on Gene Expression Analysis

Sibylle Hess and Katharina Morik
TU Dortmund University, Computer Science 8
sibylle.hess@tu-dortmund.de

High-throughput technologies permit nowadays the measurement of tens of thousands of gene expressions relating to different circumstances, such as states of health. A standard approach to summarize the relevant expressions is to cluster, i.e., to group the samples according to their gene expression profiles. However, obtaining these groups of samples poses the question: which genes are decisive to which group? This question is tackled by the task of biclustering, a simultaneous clustering of samples and genes, whose application to gene expression data is a matter of ongoing research [OKHC14]. A promising approach to solve this task is to approximate the (binarized) data matrix by a product of binary matrices, also known as Binary Matrix Factorization [ZLD$^+$10, ZDLZ07].

An apparently different task is the one Siebes et al. introduce with Krimp [VvLS11]. They invoke the Minimum Description Length (MDL) principle stated as: find the model that compresses the database best [Grü07]. Krimp returns a compact description of the dataset with a selection of patterns, i.e., subsets of features, which indicates a code table. Those are the compressing models, dictionaries of unique and prefix free codewords, which assign shorter codes to more frequently employed patterns. However, identifying the best compressing set of patterns is non-trivial. The combinatorial possibilities to define both, code table and dataset encoding, are numerous. Thus, common algorithms rely on heuristics that determine the encoding in a static order [SV12, VvLS11].

We present a unifying formulation of these tasks, pointing out how encodings describe binary matrix factorizations and vice-versa. Applying optimization methods to a relaxed formulation of this objective task enables an unrestricted usage of codes for a user-defined number of patterns. We initialize a cross-over of the applications and interpretations of the regarded data models, applying this method to the well studied medulloblastoma dataset [BTGM04]. We depict the results in terms of clustering and encoding and compare the robustness of derived models with the successfully applicable method of Nonnegative Matrix Factorization, visualizing the consistency of cluster assignments by a heatmap as proposed in [BTGM04].

References

[BTGM04] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.

[Grü07] P.D. Grünwald. *Minimum Description Length Principle.* MIT press, Cambridge, MA, 2007.

[OKHC14] Ali Oghabian, Sami Kilpinen, Sampsa Hautaniemi, and Elena Czeizler. Biclustering methods: biological relevance and application in gene expression analysis. *PloS one*, 9(3):90801, 2014.

[SV12] Koen Smets and Jilles Vreeken. Slim: Directly Mining Descriptive Patterns. In *SDM*, pages 236–247. SIAM / Omnipress, 2012.

[VvLS11] Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. Krimp: mining itemsets that compress. *Data Min. Knowl. Discov.*, 23:169–214, 2011.

[ZDLZ07] Zhongyuan Zhang, Chris Ding, Tao Li, and Xiangsun Zhang. Binary matrix factorization with applications. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 391–400. IEEE, 2007.

[ZLD$^+$10] Zhong-Yuan Zhang, Tao Li, Chris Ding, Xian-Wen Ren, and Xiang-Sun Zhang. Binary matrix factorization for analyzing gene expression data. *Data Mining and Knowledge Discovery*, 20(1):28–52, 2010.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Comparative Analysis and Phylogenetic Inference Based on Transcriptomes of 18 Chrysophyceae Species

Daniela Beisser[1], Christina Bock[2], Sabina Wodniok[2] , Nadine Nerat[2],
Jens Boenigk[2] and Sven Rahmann[1]

[1] *Genome Informatics, Institute of Human Genetics,*
University Hospital Essen, University of Duisburg-Essen, Essen, Germany
[2] *Biodiversity Department and Centre for Water and Environmental Research, University of Duisburg-Essen, Essen, Germany*

daniela.beisser@uni-due.de

High-throughput sequencing of mRNA is a promising approach to determine the expressed repertoire of genes in a species and to quantify changes in expression levels under different conditions. In this study we analyse the transcriptomes of 18 species of a widespread aquatic protistan taxon, Chrysophyceae.

For a molecular characterization of the species transcriptomes were sequenced using the Illumina platform. The sequencing reads were assembled to transcripts with Trinity [GHY+11] and annotated to KEGG orthologous genes and pathways [KG00]. In a comparative approach we investigated the differences between species as well as the influence of the trophic mode – phototrophic, mixotrophic, and heterotrophic – on gene expression and shifts in metabolic pathways.

In addition, we propose a novel transcriptome-scale approach to determine the evolutionary relationship between the organisms. Since the methods for transcriptome assemblies based on Illumina reads are still in need of improvement, an alignment-free approach is preferable over the phylogenetic inference based on multiple sequence alignments of orthologous genes. Furthermore, the calculated signature is a holistic characteristic of the expressed genome and its computation is extremely efficient.
In this study we compare and contrast both phylogenetic inference approaches using transcriptome signatures and orthologous genes identified in all 18 organisms.

References

[GHY+11] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, 29(7):644–652, Jul 2011.

[KG00] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

MeTavGen: a Taverna-based pipeline for the analysis of shotgun metagenomic data

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

Due to the decreasing costs of DNA sequencing, metagenomics became a flourishing area of research in the recent years. The amount of metagenomics datasets, their sizes and the diversity of tasks to analyse them are constantly growing. Several solutions for a semi-automatic analysis of metagenomics data are available. However, most of them can hardly be installed locally (e.g. MG-Rast, [MPD+08]; WebCarma, [GJT+09]) or are difficult to integrate with other tools (e.g. MEGAN, [HMW+11]).

Here we present MeTavGen, a flexible analysis pipeline for shotgun metagenomics datasets based on Taverna [W+13]. MeTavGen can be run locally on a single computer or an SGE computer cluster. The pipeline glues together several external software tools using custom Ruby and Bash scripts. MeTavGen can easily be extended by the user to include more analysis steps.

As input the pipeline expects either preprocessed reads or metagenomics contigs. Genes in the input sequences are annotated and their protein products are aligned to the NCBI NR database. The results are analyzed with MEGAN to determine the distribution of taxa in the metagenome. Furthermore, the genes are functionally annotated and a statistical analysis of the functional profiles of the most common taxonomic groups is performed by STAMP [PTHB14], using a similar workflow to the one adopted in [PGKL14].

The results of the analysis are presented in an automatically generated dynamic HTML report, which allows to easily access a large number of tables, metabolic pathway maps, bar and hierarchical plots, generated using MEGAN, STAMP, R and KronaTools [OBP11].

References

[GJT+09] Wolfgang Gerlach, Sebastian Jünemann, Felix Tille, Alexander Goesmann, and Jens Stoye. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC bioinformatics*, 10:430, January 2009.

[HMW+11] Daniel H Huson, Suparna Mitra, Nico Weber, Hans-Joachim Ruscheweyh, and Stephan C Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21:1552–1560, 2011.

[MPD+08] F Meyer, D Paarmann, M D'Souza, R Olson, E M Glass, M Kubal, T Paczian, a Rodriguez, R Stevens, a Wilke, J Wilkening, and R a Edwards. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9:386, January 2008.

[OBP11] Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics*, 12(1):385, January 2011.

[PGKL14] Mirjam Perner, Giorgio Gonnella, Stefan Kurtz, and Julie LaRoche. Handling temperature bursts reaching 464C: different microbial strategies in the Sisters Peak hydrothermal chimney. *Applied and environmental microbiology*, (May), May 2014.

[PTHB14] Donovan H Parks, Gene W Tyson, Philip Hugenholtz, and Robert G Beiko. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics*, 30(21):3123–3124, November 2014.

[W+13] Katherine Wolstencroft et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic acids research*, 41(Web Server issue):W557–61, July 2013.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Newtonian dynamics in the space of phylogenetic trees

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

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

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

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

References

[MV05] E. Mossel and E. Vigoda. Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees. *Sience*, 309(5744):2207–2209, 2005.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015

Circular permutations: detecting evolutionarily related protein pairs based on structural alignments

Martin Mosisch and Andrew E. Torda
Centre for Bioinformatics, University Hamburg
mosisch@zbh.uni-hamburg.de

Circularly permutated proteins are an event in which part of the C-terminus has moved to the N-terminus or vice versa, as if on a circle. Thus a protein ABCD may be related to BCDA or DABC. These permutations are thought to arise from gene duplications/deletions, fusions and fissions or the cutting and pasting mechanisms of the restriction modification system. [WBB06]

Such evolutionary events are considered to be rare, but their detection can help determining structural domains and engineering recombinant proteins. [Kam11]

With traditional dynamic programming methods, these rearrangements lead to partial similarities, meaning that distant evolutionary events will be missed. We use an extended alignment matrix to detect permuted relations, but perform structural instead of sequence comparisons. [MST09]

Our method is not free of thresholds and adjustable parameters, but as set now, we find 95% of previously documented examples. We can, however, process very large data sets, while still working with structural alignments. Applying the method to a set of about 18 × 10^6 candidate pairs leads to many surprises. Circular permutations are not as infrequent as one might expect and about 85 % of the detected events correspond to annotated protein domain boundaries [VBAS04][MBDG+15].

References

[Kam11] M. Kamionka. Engineering of Therapeutic Proteins Production in Escherichia coli. *Curr Pharm Biotechno*, 12:268–274, 2011.

[MBDG$^+$15] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, and M. Gwadz et al. CDD: NCBI's Conserved Domain Database. *Nucleic Acids Res*, 43:D222–D226, 2015.

[MST09] T. Margraf, G. Schenk, and A. E. Torda. The SALAMI Protein Structure Search Server. *Nucleic Acids Res*, 37:W480–W484, 2009.

[VBAS04] S. Veretnik, P. E. Bourne, N. N. Alexandrov, and I. N. Shindyalov. Toward Consistent Assignment of Structural Domains in Proteins. *J Mol Biol*, 339:647–678, 2004.

[WBB06] J. Weiner and E. Bornberg-Bauer. Evolution of Circular Permutations in Multidomain Proteins. *Mol Biol Evol*, 23:734–743, 2006.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1350v2 | CC-BY 4.0 Open Access | rec: 17 Sep 2015, publ: 17 Sep 2015