Invited presentations, junior research groups and research highlights at GCB 2015

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

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

EDITORIAL

There are several types of presentations at the German Conference on Bioinformatics (GCB): original research (proceedings track), highlights, posters, and, for the first time in 2015, junior research group presentations.
This document collects the abstracts of invited presentations, of the junior research groups and of research highlights from the past year that were presented at GCB 2015. The program committee accepted 8 out of 10 submitted highlight abstracts and 6 out of 8 submitted junior research group presentations.
As junior research group presentations are new to the conference, we would especially invite feedback about this new track, in addition to other feedback you may have about the conference.

Keywords: GCB 2015, abstracts, invited talks, junior research groups, highlights

PeerJ PrePrints

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

PeerJ PrePrints

Invited Talks

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

Modelling Coverage in RNA Sequencing

Arndt von Haeseler
Max F. Perutz Laboratories, Center for Integrative Bioinformatics Vienna (CIBIV)
http://www.cibiv.at/~haeseler/

RNA sequencing (RNA-seq) is the method of choice for measuring the expression of RNAs in a cell population. In an RNA-seq experiment, sequencing the full length of larger RNA molecules requires fragmentation into smaller pieces to be compatible with limited read lengths of most deep-sequencing technologies. Unfortunately, the issue of non-uniform coverage across a genomic feature has been a concern in RNA-seq and is attributed to preferences for certain fragments in steps of library preparation and sequencing. However, the disparity between the observed non-uniformity of read coverage in RNA-seq data and the assumption of expected uniformity elicits a query on the read coverage profile one should expect across a transcript, if there are no biases in the sequencing protocol. We propose a simple model of unbiased fragmentation where we find that the expected coverage profile is not uniform and, in fact, depends on the ratio of fragment length to transcript length. To compare the non-uniformity proposed by our model with experimental data, we extended this simple model to incorporate empirical attributes matching that of the sequenced transcript in an RNA-seq experiment. In addition, we imposed an experimentally derived distribution on the frequency at which fragment lengths occur.

We used this model to compare our theoretical prediction with experimental data and with the uniform coverage model. If time permits, we will also discuss a potential application of our model.

This is joint work with Celine Prakash, Florian Pflug and Luis Felipe Paulin Paz.

PeerJ PrePrints

LoRDEC: a tool for correcting errors in long sequencing reads

Eric Rivals
Laboratoire d'Informatique de Microélectronique et de Robotique de Montpellier (LIRMM) and Institute of Computational Biology (IBC) CNRS and Université de Montpellier, France
http://www.lirmm.fr/~rivals/

High-throughput DNA/RNA sequencing is a routine experiment in molecular biology and life sciences in general. For instance, it is increasingly used in the hospital as a key procedure of personalized medicine. Compared to the second generation, third generation sequencing technologies produce longer reads with comparatively lower throughput and higher error rate. Those errors include substitutions, indels, and they hinder or at least complicate downstream analysis like mapping or de novo assembly. However, these long read data are often used in conjunction with short reads of the 2nd generation.

I will present a hybrid strategy for correcting the long reads using the short reads that we introduced last year. Unlike existing error correction tools, ours, called LoRDEC, avoids aligning short reads on long reads, which is computationally intensive. Instead, it takes advantage of a succinct graph to represent the short reads, and compares long reads to paths in the graph. Experiments show that LoRDEC outperforms existing methods in running time and memory while achieving a comparable correction performance. It can correct both Pacific Biosciences and MinION reads from Oxford Nanopore.

LoRDEC is available at http://atgc.lirmm.fr/lordec. This is joint work with L. Salmela and A. Makrini.

PeerJ PrePrints

From sequence analysis to graph analysis

Veli Mäkinen
Department of Computer Science, University of Helsinki
http://www.cs.helsinki.fi/u/vmakinen/

The abstraction of a genome as a linear sequence has created a vast sequence analysis literature with plethora of interesting subproblems defined and often algorithmically optimally solved; recent results in compressed indexing provide linear time sequence analysis functionality even in space close to what an input sequence occupies. One could say it is time to move on to more realistic abstractions of genomic content. This talk explores what happens to a selected classical sequence analysis tasks when labeled directed acyclic graphs (labeled DAGs) are used as inputs. Applications in partially phased diploid genomes, pan-genomes, and splicing graphs, are discussed. Some algorithms for the new problems are presented. The talk concludes with a list of open problems to summarize what needs to be achieved in order for the theory of labeled DAG analysis to reach completion similar to sequence analysis.

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

ELIXIR Europe: the European life science infrastructure for biological data

Andrew Smith
EMBL-EBI / ELIXIR Europe
https://www.elixir-europe.org/

The life sciences are undergoing a transformation. Scientists are rapidly generating the most complex and heterogeneous datasets that science can currently imagine, with unprecedented volumes of biological data to manage. Data will only generate long-term value if it is Findable, Accessible, Interoperable and Re-usable ('FAIR'). This requires a scalable infrastructure that connects local, national and European efforts and provides standards, tools and training for data management and analysis.

Established in January 2014, ELIXIR - the European life science Infrastructure for Biological Information - is a distributed organisation comprising national bioinformatics research infrastructures across Europe and the European Bioinformatics Institute (EMBL-EBI). This coordinated infrastructure supports data standards, exchange, interoperability, storage, security and training. From September 2015, the newly-awarded ELIXIR-EXCELERATE Horizon 2020 grant will fast-track ELIXIR's early implementation phase by coordinating and enhancing existing resources into a world-leading data service for academia and industry and growing bioinformatics capacity and competence across Europe.

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

Intra-tumour heterogeneity and genomic rearrangements in human malignancies

Roland Schwarz
EMBL-EBI
https://www.ebi.ac.uk/about/people/roland-schwarz

Accurate reconstruction of the evolutionary history of cancer in the patient and quantification of intra-tumour heterogeneity (ITH) are current challenges in cancer genomics. Genomic rearrangements are thereby of particular importance, but notoriously difficult to deal with computationally. The accuracy of tree inference from genomic rearrangements further depends on the quality of the phasing of copy-numbers: the assignment of major and minor copy-numbers to the two physical parental alleles. So far, phasing has been done using evolutionary criteria alone, a heuristic and computationally expensive procedure which impedes probe-level resolution tree reconstruction.

I will give an overview of the challenges and current state of research in reconstructing cancer trees from copy-number data. Results from our clinical studies demonstrate how ITH is associated with chemotherapy resistance in the clinic. I will further illustrate the importance of haplotype-specific copy-number assignment and show how the common genetic background between multiple samples from the same patient can be used to accurately phase copy-number data. This is a crucial step towards probe-level resolution tree inference on genomic rearrangement events in cancer and exact quantification of genetic heterogeneity for routine applications in translational cancer research.

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

PeerJ PrePrints

Junior Research Groups

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

Virus-Host Transcriptomics

Caroline C. Friedel
Institut für Informatik, Ludwig-Maximilians-Universität München
caroline.friedel@bio.ifi.lmu.de

PeerJ PrePrints

Introduction

The application of next generation sequencing technologies (NGS) to sequencing of RNA (RNA-seq) provides novel opportunities for the analysis of transcriptomes beyond the simple quantification of gene expression. In particular, the combination of RNA-seq with powerful techniques for selecting specific types of RNA (e.g. newly transcribed RNA using 4sU-tagging [DRR+08] or actively translated RNA using ribosome profiling [IGNW09]) now allows quantification of real-time changes in RNA synthesis [MLW+12], RNA processing [WBB+12], and translation [IGNW09].

A further interesting application arises from the fact that RNA-seq protocols do not distinguish between RNA from different species. Thus, in case of infections by viruses or bacteria, RNA from the infecting species will automatically be sequenced together with the host RNA. Originally, this application has been proposed in a thought experiment by Westermann et al. [WGV12] and denoted as dual RNA-seq, although it is not limited to just one infecting species and the host. For instance, Castellarin *et al.* [CWF+12] identified a number of microbes in RNA-seq data of colorectal carcinoma and normal tissue samples. To date, dual RNA-seq has been used to annotate and quantify the transcriptome and translatome of several herpesviruses, which are large DNA viruses that replicate in the nucleus. This includes murine and human cytomegalovirus (MCMV and HCMV) [MLW+12, SGWM+12], Kaposi's sarcoma-associated herpesvirus (KSHV) [AWSG+14], and human herpesvirus 1 (HSV-1) [REL+15].

In this presentation, I will provide an overview on methods developed in my group for the analysis of RNA-seq data of infected cells, in particular for the analysis of transcriptional and translational activity, transcription termination and RNA processing during lytic HSV-1 infection [REL+15]. This includes methods for parallel RNA-seq mapping against several read sources [BCZF12, BCZF13, BKC+15] as well as quantification of transcription termination and polyadenylation sites in both host and virus.

Parallel RNA-seq mapping to virus and host

One major challenge in both "standard" and dual RNA-seq is the identification of the transcriptomic origin of sequencing reads (mapping). Accordingly, a number of software programs have been developed for this task, e.g. TopHat [TPS09] or STAR [DDS+13]. However, these approaches to not directly support mapping of reads from multiple species or other read sources (e.g. rRNA sequences, which are not included in the human reference genome). Although additional sequences may be included into the mapping index, this either requires reindexing all reference sequences including the host genome for each new virus investigated or always mapping against all microbe and virus genomes. In addition, non-unique alignments are generally not resolved, which is a problem for rRNA reads which also map to rRNA pseudogenes in the host genome or a meta-transcriptomic screen against all known microbe and virus genomes. To address this problem, we recently extended our context-based RNA-seq mapping approach ContextMap [BCZF12] to allow parallel mapping against different read sources resulting in a unique mapping of each read to only one species/read source [BCZF13].

The parallel mapping approach could be integrated easily into ContextMap as even in the original implementation initial read alignments are clustered into so-called contexts that are treated independently until the last integrating step. Essentially a context represents a set of reads originating from the same

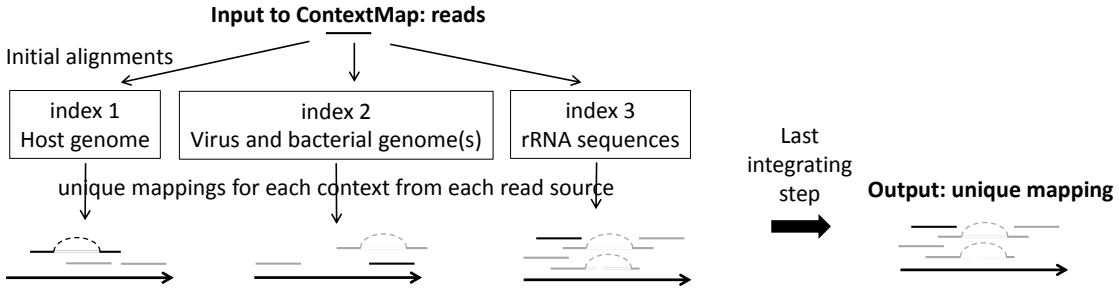


Figure 1: Parallel mapping against host, virus and bacterial genomes as well as rRNA is realized in ContextMap by (1) performing initial alignments against indices for several species/read sources to define contexts, (2) identifying best alignments for each read independently within each context and (3) resolving the resulting multiple alignments in the final integrating step.

PeerJ PrePrints

stretch of the genome and likely corresponding to transcripts of the same or overlapping genes. Multiple alignments of reads to different contexts are allowed, which are then resolved in the last step. Thus, parallel mapping to multiple species could be included in ContextMap in a straightforward way by aligning against multiple sequence indices in the initial alignment step to recover contexts for different species (Figure 1). This approach is also included in the recent ContextMap 2 release, which allows the use of alternative short read alignment programs and can recover reads containing multiple exon-exon junctions or insertions or deletions [BKC+15].

Wide-spread disruption of host transcription termination in HSV-1

Parallel analysis of host and virus transcription and translation can lead to highly interesting insights not only into the infection process itself but also into important biological processes. This was illustrated by our recent study on HSV-1 lytic infection [REL+15], which established HSV-1 infection as an interesting model system to study transcription termination. HSV-1 is an important human pathogen that causes both common cold sores as well as life-threatening infections and rapidly shuts down host gene expression during lytic infection. In our study, we combined sequencing of 4-thiouridine (4sU)-labeled newly transcribed RNA (4sU-RNA) and ribosome profiling to study both host and virus transcription and translation during the full course of HSV-1 lytic infection. 4sU-labeling was performed in one-hour intervals during the first 8 hours of infection and ribosome profiling was performed at 0, 1, 2, 4, 6 and 8h post infection (p.i.).

Surprisingly, we found that the transcriptional up-regulation of 659 cellular genes was not matched by a respective increase in translational activity. Only 33 (0.34%) of translated genes showed increased translational activity at 8h p.i. When analyzing genes that were transcriptionally induced but not translated, we observed massive transcriptional activity upstream of their 5'-ends at late times of infection originating from neighboring upstream genes (Figure 2). This suggested that the transcription termination and cleavage machinery did no longer recognize or properly function at the termination signals of upstream genes, resulting in transcription into downstream regions by >100,000nt (denoted as 'read-out'). We found that read-out affected the majority of cellular genes and was correlated with a higher prevalence of non-canconical polyadenylation [poly(A)] signals. Although this indicated that non-canonical and likely weaker poly(A) signals were more strongly affected by disrupted transcription termination, the majority of genes with read-out still had the canonical AAUAAA poly(A) signal. Thus, poly(A) signal strength is certainly not the only factor influencing the extent of read-out.

Late in infection, read-out commonly extended over thousands of nucleotides into downstream genes (denoted as 'read-in'). At least 32% of genes showed >15% read-in at 8h p.i. and the extent of read-in depended on the distance to the next upstream gene. For genes with low or no transcription in uninfected

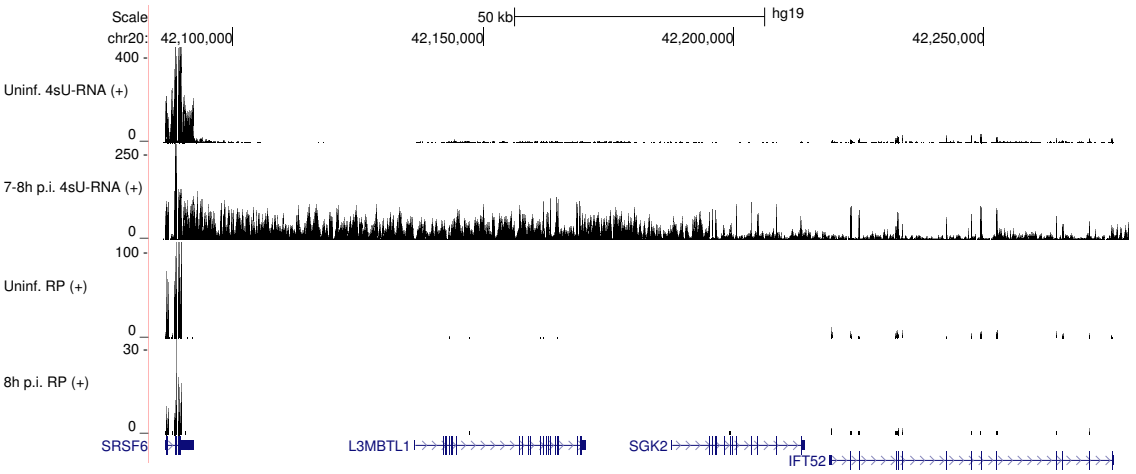


Figure 2: Disruption of transcription termination of the SRSF6 gene and read-in into the downstream SGK2 and IFT52 genes. The top two rows show transcriptional activity (4sU-RNA) and the bottom two rows translational activity (ribosome profiling, RP) in uninfected cells and at 8h p.i., respectively.

cells, read-in often exceeded endogenous transcript level, resulting in seeming 'induction'. This explained the discrepancy between transcriptional and translational induction. Furthermore, HSV-1 infection induced aberrant splicing events, which were enriched among genes with high read-out. Thus, splicing was already affected upstream of poly(A) sites suffering from read-out. Interestingly, 44% of the induced splice junctions were novel and 11% of these represented intergenic splicing between two neighboring genes connected by read-out and subsequent read-in. These intergenic splicing events thus conclusively demonstrated that disruption of transcription termination resulted in large RNA molecules spanning two or more cellular genes.

To investigate whether disruption of transcription termination was specific to the host or also affected HSV-1 genes, we identified reads containing part of a poly(A) tail, i.e. reads for which a partial alignment of the read start to the host or HSV-1 genome was followed by a stretch of A's. As coverage of the poly(A) tails was generally at least two orders of magnitudes lower than of the corresponding transcripts, only few poly(A) reads were recovered for the host genome. Coverage of HSV-1 transcripts, however, was in the order of tens-of-thousands of reads per genome position, allowing us to quantify poly(A) site usage of all but one viral gene (see Figure 3 for the UL39-50 gene segment). Viral poly(A) sites were almost exclusively preceded by an AAUAAA poly(A) signal. To investigate changes in poly(A) site usage in the whole HSV-1 genome throughout infection, we correlated gene expression upstream of each poly(A) site with the number of identified poly(A)-tailed reads. For 80% of poly(A) sites, this correlation was >0.9, which argued against regulated poly(A) site usage in HSV-1 infection and showed that disruption of transcription termination was host-specific.

References

[AWSG+14] Carolina Arias, Ben Weisburd, Noam Stern-Ginossar, Alexandre Mercier, Alexis S. Madrid, Priya Bellare, Meghan Holdorf, Jonathan S. Weissman, and Don Ganem. KSHV 2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features. *PLoS Pathog*, 10(1):e1003847, Jan 2014.

[BCZF12] Thomas Bonfert, Gergely Csaba, Ralf Zimmer, and Caroline C. Friedel. A context-based approach to identify the most likely mapping for RNA-seq experiments. *BMC Bioinformatics*, 13 Suppl 6:S9, 2012.

[BCZF13] Thomas Bonfert, Gergely Csaba, Ralf Zimmer, and Caroline C. Friedel. Mining RNA-seq data for infections and contaminations. *PLoS One*, 8(9):e73071, 2013.

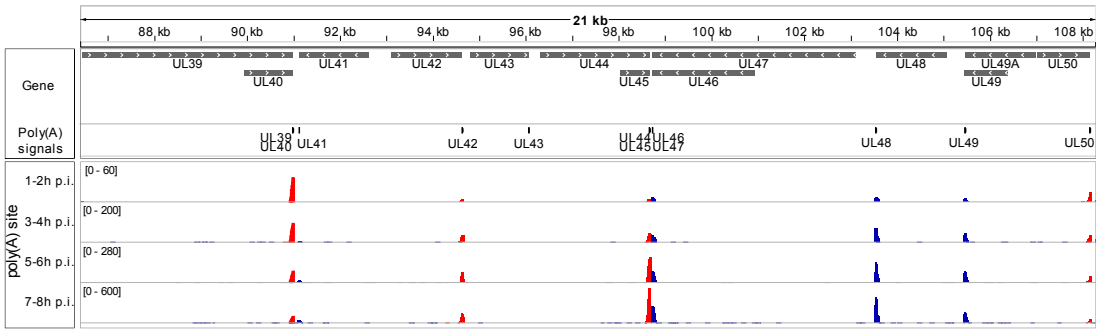


Figure 3: Poly(A) tail read coverage in 4sU-RNA for the UL39-50 gene segment (red = positive strand, blue = negative strand).

[BKC+15] Thomas Bonfert, Evelyn Kirner, Gergely Csaba, Ralf Zimmer, and Caroline C. Friedel. ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinformatics*, 16(1):122, 2015.

[CWF+12] Mauro Castellarin, Ren L. Warren, J Douglas Freeman, Lisa Dreolini, Martin Krzywinski, Jaclyn Strauss, Rebecca Barnes, Peter Watson, Emma Allen-Vercoe, Richard A. Moore, and Robert A. Holt. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res*, 22(2):299–306, Feb 2012.

[DDS+13] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[DRR+08] Lars Dölken, Zsolt Ruzsics, Bernd Rädle, Caroline C. Friedel, Ralf Zimmer, Jörg Mages, Reinhard Hoffmann, Paul Dickinson, Thorsten Forster, Peter Ghazal, and Ulrich H. Koszinowski. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, 14(9):1959–1972, Sep 2008.

[IGNW09] Nicholas T. Ingolia, Sina Ghaemmaghami, John R S. Newman, and Jonathan S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, Apr 2009.

[MLW+12] Lisa Marcinowski, Michael Lidschreiber, Lukas Windhager, Martina Rieder, Jens B. Bosse, Bernd Rädle, Thomas Bonfert, Ildiko Gyry, Miranda de Graaf, Olivia Prazeres da Costa, Philip Rosenstiel, Caroline C. Friedel, Ralf Zimmer, Zsolt Ruzsics, and Lars Dölken. Real-time transcriptional profiling of cellular and viral gene expression during lytic cytomegalovirus infection. *PLoS Pathog*, 8(9):e1002908, Sep 2012.

[REL+15] Andrzej J. Rutkowski, Florian Erhard, Anne L'Hernault, Thomas Bonfert, Markus Schilhabel, Colin Crump, Philip Rosenstiel, Stacey Efstathiou, Ralf Zimmer, Caroline C. Friedel, and Lars Dölken. Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun*, 6:7126, 2015.

[SGWM+12] Noam Stern-Ginossar, Ben Weisburd, Annette Michalski, Vu Thuy Khanh Le, Marco Y. Hein, Sheng-Xiong Huang, Ming Ma, Ben Shen, Shu-Bing Qian, Hartmut Hengel, Matthias Mann, Nicholas T. Ingolia, and Jonathan S. Weissman. Decoding human cytomegalovirus. *Science*, 338(6110):1088–1093, Nov 2012.

[TPS09] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[WBB+12] Lukas Windhager, Thomas Bonfert, Kaspar Burger, Zsolt Ruzsics, Stefan Krebs, Stefanie Kaufmann, Georg Malterer, Anne L'Hernault, Markus Schilhabel, Stefan Schreiber, Philip Rosenstiel, Ralf Zimmer, Dirk Eick, Caroline C. Friedel, and Lars Dölken. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res*, 22(10):2031–2042, Oct 2012.

[WGV12] Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol*, 10(9):618–630, Sep 2012.

PeerJ PrePrints

Development and application of computational methods for the identification and optimization of bioactive compounds

Johannes Kirchmair
Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany
kirchmair@zbh.uni-hamburg.de

In September 2014 we started building up a new research lab for applied cheminformatics and molecular design at the Center for Bioinformatics of the University of Hamburg. Our major research interests are the development and application of computational methods for the identification and optimization of bioactive compounds, in particular for drug metabolism prediction, target prediction, virtual screening, conformer ensemble generation, and natural product research. Here we provide a brief summary of our recent activities in two of these research areas.

PeerJ PrePrints

1 Development of computational methods for drug metabolism prediction

Metabolism of small organic molecules can yield metabolites with substantially different physicochemical and biological properties [Kirchmair 2013a]. Consequently, understanding biotransformation is of immediate relevance to the safety and efficacy of drugs, cosmetics, nutritional supplements and agrochemicals. Today a plethora of experimental methods are available which allow the generation of a fairly complete picture of the metabolic fate of small molecules but they remain expensive and time-consuming. Driven by these factors as well as ethical issues related to the use of animal models, computational methods for drug metabolism prediction have become an active field of research [Olsen 2015, Peach 2012, Kirchmair 2015, Kirchmair 2014, Kirchmair 2012].

The primary bottleneck for computational tools is the scarcity of high-quality data on drug metabolism. The largest database on metabolic reactions, Metabolite (BIOVIA, San Diego, CA), contains about 100k substrate-metabolite pairs organised in about 14k metabolic pathways. It covers about 90% of all approved drugs but only a small fraction of drug-like molecules, natural products and human endogenous metabolites.

We have developed data mining methods for the automated analysis of metabolic reactions and pathways such as the ones stored in Metabolite. In a first study we used these techniques to analyse how and to which extent the metabolic system changes the physicochemical properties of small organic molecules (including drugs, endogenous metabolites, and molecules related to traditional Chinese medicine) [Kirchmair 2013a]. For example, we could show that drug metabolism produces metabolites with a calculated logP that is on average one log unit lower than that of the parent compound. Interestingly, this shift toward more hydrophilic molecules is much less pronounced for endogenous metabolites such as nutrients and micronutrients, which are retained in the body. Such methods allowed us to identify specific metabolic reactions and enzymes, which, against the global trend, result in more lipophilic metabolites. This knowledge e.g. can be applied to the design of skin care products with a prolonged retention time at the target tissue.

In a second study we developed a random forest-based predictor of sites of metabolism (regioselectivity): FAME (FAst MEtabolizer) [Kirchmair 2013b]. The sites of metabolism of about 20k molecules of Metabolite were automatically annotated using the MetaPrint2D software framework [Adams 2010]. Seven 2D chemical descriptors encoding the element, hybridisation state, electronic properties and steric accessibility were calculated for all atoms of all molecules. A collection of random forest models was then trained on subsets of this data. Individual models were computed for human, rat, dog and mammalian metabolism. Reaction classification was used to derive dedicated models for phase 1 and phase 2 metabolism. FAME correctly identifies at least one known site of metabolism among the top-

1, top-2, and top-3 highest-ranked atom positions in up to 71%, 81%, and 87% of all cases tested, respectively. These success rates are comparable to or better than other models focused on specific enzyme families (such as cytochrome P450s; CYPs), yet FAME covers a very broad chemical space (drugs, endogenous metabolites and natural products) and a fairly comprehensive set of reactions and enzymes relevant to xenobiotic metabolism. In a complementary approach we used three probabilistic machine learning methods, Parzen-Rosenblatt Window (PRW), Naive Bayesian (NB) and RASCAL (Random Attribute Subsampling Classification ALgorithm) for the generation of highly accurate models for site of metabolism prediction [Tyzack 2014]. The classifiers were implemented in CUDA/C++ for GPU acceleration and obtained top-2 success rates of about 80-90% for CYPs 3A4, 2D6 and 2D9.

A plateauing in prediction accuracy of methods for site of metabolism prediction is observed, and the primary reason for this appears to be the limitations of current datasets with respect to coverage of the chemical space, diversity, completeness, correct assignment of sites of metabolism and stereochemical information. Here, substantial efforts in data collection and curation are to be made. Future research directions will include the implementation of more advanced descriptors for a more accurate representation of the chemical reactivity and steric accessibility of atoms.

2 Computer-guided identification and optimization of bioactive compounds

Currently our research group is actively pursuing a dozen national and international research collaborations with experimentalists on the identification and optimization of bioactive compounds. Much of our recent drug discovery projects have been focussed on viral [Richter 2015, Grienke 2011, Kirchmair 2011, Grienke 2010, von Grafenstein 2015] and bacterial [von Grafenstein 2015, Walther 2015] neuraminidases, which are scientifically highly interesting to address using computational techniques because of their pronounced conformational flexibility and specific structural properties. For example, influenza virus neuraminidase is only active when in a quaternary assembly, but neuraminidases of other biological systems, such as bacteria, are active as monomers. Using molecular dynamics simulation techniques we derived a hypothesis of the structural basis of this assembly dependency [von Grafenstein 2015]. Understanding this specific requirement of influenza neuraminidases is of immediate relevance to the structure-based design of new inhibitors, which so far has often relied on structures derived from simulations of the monomer of the viral enzyme. As we know now, simulation of this system is most likely insufficient because of the significant impact of the assembly state on the conformation of the active site.

We successfully applied a shape-based screening method to identify a variety of plant constituents from *Glycyrrhiza glabra* [Grienke 2011] and others [Kirchmair 2011] as inhibitors of influenza neuraminidase. A shape-based screening method also allowed us to identify benzylhydantoin and related compounds as *in vivo* highly effective chemical chaperons of phenylalanine hydroxylase [Santos-Sierra 2012]. These compounds can be used to treat phenylketonuria, an inherited deficiency caused by protein misfolding. Current treatment options for this disease are very limited, costly and often not effective.

Recently we identified inhibitors of the interaction of protein kinase C (PKC) epsilon and RACK2 [Rechfeld 2015] using a pharmacophore model. PKCepsilon has been related to neoplastic transformation, cardiac hypertrophy, nociceptor function and others. The model was derived from the structure of the C2 domain of PKCepsilon and used to screen a commercial library of 330k molecules for potential disruptors. Nineteen compounds were purchased and tested in *in vitro* assays. One of the tested compounds (based on a thienoquinoline scaffold) showed moderate activity as a disruptor of this protein-protein interaction, and the best out of 19 analogues tested in a follow-up study, N-(3-acetylphenyl)-9-amino-2,3-dihydro-1,4-dioxino[2,3-g]thieno[2,3-b]quinoline-8-carboxamide, had an IC_{50} of 5.9 micromolar (which is comparable to that of a dodecapeptide fragment of RACK2 binding to this protein-protein interface).

PeerJ PrePrints

References

[Adams 2010] Adams, S. E. Molecular similarity and xenobiotic metabolism. Ph.D. Thesis. University of Cambridge, UK. 2010.

[Grienke 2011] Grienke, U.; Braun, H.; Seidel, N.; Kirchmair, J.; Richter, M.; Krumbholz, A.; von Grafenstein, S.; Liedl, K. R.; Schmidtke, M.; Rollinger, J. M. Computer-guided approach to access the anti-influenza activity of licorice constituents. *Journal of Natural Products*, 2014, 77, 563-570.

[Grienke 2010] Grienke, U.; Schmidtke, M.; Kirchmair, J.; Pfarr, K.; Wutzler, P.; Drrwald, R.; Wolber, G.; Liedl, K. R.; Stuppner, H.; Rollinger, J. M. Antiviral potential and molecular insight into neuraminidase inhibiting diarylheptanoids from *Alpinia katsumadai*. *Journal of Medicinal Chemistry*, 2010, 53, 778-786.

[Kirchmair 2015] Kirchmair, J.; Gller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting drug metabolism: Experiment and/or computation? *Nature Reviews Drug Discovery*, 2015, 14, 387-404.

[Kirchmair 2014] Kirchmair, J. (ed.) Drug Metabolism Prediction Wiley: Weinheim, 2014. ISBN 978-3-527-33566-4

[Kirchmair 2013a] Kirchmair, J.; Howlett, A.; Peironcely, J.; Murrell, D. S.; Williamson, M. J.; Adams, S. E.; Hankemeier, T.; van Buren, L.; Duchateau, G.; Klaffke, W.; Glen, R. C. How do metabolites differ from their parent molecules and how are they excreted? *Journal of Chemical Information and Modeling*, 2013, 53, 354-367.

[Kirchmair 2013b] Kirchmair, J.; Williamson, M. J.; Afzal, A. M.; Tyzack, J. D.; Choy, A. P. K.; Howlett, A.; Rydberg, P.; Glen, R. C. FAst MEtabolizer (FAME): A rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes. *Journal of Chemical Information and Modeling*, 2013, 53, 2896-2907.

[Kirchmair 2012] Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational prediction of metabolism: Sites, products, SAR, P450 enzyme dynamics and mechanisms. *Journal of Chemical Information and Modeling*, 2012, 52, 617-648.

[Kirchmair 2011] Kirchmair, J.; Rollinger, J. M.; Liedl, K. R.; Seidel, N.; Krumbholz, A.; Schmidtke, M. Novel neuraminidase inhibitors: Identification, biological evaluation and investigations of the binding mode. *Future Medicinal Chemistry*, 2011, 3, 437-450.

[Olsen 2015] Olsen, L.; Oostenbrink, C.; Jrgensen, F. S. Prediction of cytochrome P450 mediated metabolism. *Advanced Drug Delivery Reviews*, 2015, doi: 10.1016/j.addr.2015.04.020.

[Peach 2012] Peach, M. L.; Zakharov, A. V.; Liu, R.; Pugliese, A.; Tawa, G.; Wallqvist, A.; Nicklaus, M.C. Computational tools and resources for metabolism-related property predictions. 1. Overview of publicly available (free and commercial) databases and software. *Future Medicinal Chemistry*, 2012, 4, 1907-1932.

[Rechfeld 2015] Rechfeld, F.; Gruber, P.; Kirchmair, J.; Boehler, M.; Hauser, N.; Hechenberger, G.; Garczarczyk, D.; Lapa, G. B.; Preobrazhenskaya, M. N.; Goekjian, P.; Langer, T.; Hofmann, J. Thienoquinolines as novel disruptors of the PKC/RACK2 protein-protein interaction. *Journal of Medicinal Chemistry*, 2014, 57, 3235-3246.

[Richter 2015] Richter, M.; Schumann, L.; Walther, E.; Hoffmann, A.; Braun, H.; Grienke, U.; Rollinger, J. M.; von Grafenstein, S.; Liedl, K. R.; Kirchmair, J.; Wutzler, P.; Sauerbrei, A.; Schmidtke, M. Complementary assays helping to overcome challenges for identifying neuraminidase inhibitors. *Future Virology*, 2015, 10, 77-88.

[Santos-Sierra 2012] Santos-Sierra, S.; Kirchmair, J.; Perna, A. M.; Rei, D.; Kemter, K.; Rschinger, W.; Glossmann, H.; Gersting, S. W.; Muntau, A. C.; Wolber, G.; Lagler, F. Novel pharmacological chaperones that correct phenylketonuria in mice. *Human Molecular Genetics*, 2012, 21, 1877-1887.

[Tyzack 2014] Tyzack, J. D.; Mussa, H. Y.; Williamson, M. J.; Kirchmair, J.; Glen, R. C. Cytochrome P450 site of metabolism prediction from 2D topological fingerprints using GPU accelerated probabilistic classifiers. *Journal of Cheminformatics*, 2014, 29.

[von Grafenstein 2015] von Grafenstein, S.; Wallnfer, H. G.; Kirchmair, J.; Fuchs, J. E.; Huber, R. G.; Spitzer, G.; Schmidtke, M.; Rollinger, J. M.; Liedl, K. R. Interface dynamics explain assembly dependency of influenza neuraminidase catalytic activity. *Journal of Biomolecular Structure & Dynamics*, 2015, 33, 104-120.

[Walther 2015] Walther, E.; Richter, M.; Xu, Z.; Kramer, C.; von Grafenstein, S.; Kirchmair, J.; Grienke, U.; Rollinger, J. M.; Liedl, K. R.; Slevogt, H.; Sauerbrei, A.; Saluz, H. P.; Pfister, W.; Schmidtke, M. Antipneumococcal activity of neuraminidase inhibiting artocarpin. *International Journal of Medical Microbiology*, 2015, 305, 289-297.

PeerJ PrePrints

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

Algorithms for Computational Genomics

Tobias Marschall
Center for Bioinformatics, Saarland University, Saarbrücken, Germany
Max-Planck-Institute for Informatics, Saarbrücken, Germany
t.marschall@mpi-inf.mpg.de

Abstract: The topics studied in the Algorithms for Computational Genomics group range from theoretical foundations in algorithmic statistics, combinatorial optimization, and sequence algorithms to applied studies on population genetics, structural variation in human, and horizontal gene transfer in bacteria. We aim to develop algorithmic concepts as well as to provide production quality software tools. Current topics addressed in the group include structural variation calling and genotyping, read-based phasing of diploid individuals and viral quasispecies, methods for detecting horizontal gene transfer, as well as computational pan-genomics.

PeerJ PrePrints

1 Group Development

The *Algorithms for Computational Genomics* group was established in April 2014, when Tobias Marschall was appointed assistant professor ("Juniorprofessor") at the Center for Bioinformatics at Saarland University. Since then, the group is also affiliated with the Max-Planck-Institute for Informatics where Tobias has been appointed Senior Researcher. Two PhD students (Shilpa Garg and Ali Ghaffaari) joined the group in November 2014 and April 2015, respectively. Furthermore, five Master and six Bachelor students are members of the group, working on their respective thesis projects.

2 Research Strategy

The group develops algorithms and statistical methods for computational genomics. In particular, we work on methods to analyze high-throughput sequencing data to study genetic diversity in human populations, bacterial adaptation, and cancer. On the one hand, we develop the required theoretical foundations in algorithmic statistics, combinatorial optimization, and sequence algorithms and, on the other hand, we apply the resulting methods in collaboration with biomedical researchers to gain biological insights in the aforementioned domains.

3 Research Areas

Topics addressed in the group range from algorithms for low level data processing to questions of population genetics. At present, we are particularly focusing on the following projects.

3.1 Structural Variation Calling and Genotyping

Beyond SNPs and short indels, larger genetic differences between individuals make an important contribution to genetic diversity in human populations [KUA+07, MWS+11]. Such larger events, called *structural variants (SVs)*, come in the form of deletions, insertions, duplications, translocations, inversions, and also more complex events. Detecting SVs from next-generation sequencing (NGS) data has been subject to active research, as reviewed in [MSB09] and [ACE11]. While still a postdoc at CWI Amsterdam, Tobias Marschall developed CLEVER [MCC+12] and MATE-CLEVER [MHS13], two ap-

proaches to detect deletions and insertions. The main contribution of these methods was to achieve good performance also for the particularly difficult mid-size deletions between 30–250bp (called deletion twilight zone by some). MATE-CLEVER also introduced a novel Bayesian approach for Mendelian-inheritance-aware genotyping of insertions and deletions. In a current project, we show that extending these approaches to inversions and duplications yields performance clearly superior to existing genotyping methods (yet unpublished).

In a recent publication [LMP+15], we furthermore contributed to establishing a virtual-machine based platform for benchmarking and running a multitude of SV calling algorithms. This helps to alleviate practical problems like (missing) software dependencies or incompatible data formats and, more importantly, facilitates reliable and reproducible research.

Beyond such practical problems, more fundamental issues exist regarding the seemingly simple task of comparing or merging multiple sets of SV calls. The interplay of two effects renders this a non-trivial task: on the one hand, SV callers in general do not deliver single-base-pair resolution and, on the other hand, two SVs with different breakpoint coordinates can be equivalent in the presence of repeats (in the sense that the resulting donor sequences are identical). We recently introduced a framework addressing both aspects simultaneously and provided an efficient implementation [WMSM15].

3.2 Structural Variations in the Genome of the Netherlands.

The Genome of the Netherlands (GoNL) project has sequenced the whole genomes of 750 Dutch individuals from 250 families. Applications of these data include building high-quality reference panels for imputation, studying *de novo* mutations and the corresponding mechanisms, estimating the rate of such events, and analyzing population structure, among many others. We contributed to this project [The14] as part of the Structural Variations subgroup and provided algorithms for the discovery and genotyping of structural variations, especially for "difficult" types like mid-size deletions and insertions. Furthermore, we addressed the particularly challenging task of detecting *de novo* SVs, i.e. structural variants present in a child and *not* inherited from any parent, published as [KFH+15]. Presently, we work on phasing and imputation of structural variations found in the GoNL.

3.3 Haplotype Reconstruction—Diploid Case.

Reconstructing the two haplotypes of a diploid organism (also known as phasing) is an important problem with applications in fundamental research but also in clinical settings, as discussed in [GCR14]. Emerging sequencing technologies hold the promise of allowing for read-based phasing through longer reads. On the computational side, most formalizations of the corresponding optimization problem are NP-hard. In an approach called WhatsHap [PMP+15], we demonstrated that (i) the problem instances encountered in practice can be solved using a fixed parameter tractable (FPT) algorithm and (ii) that read-based phasing indeed delivers excellent performance for long reads. In follow-up work, we contributed to an optimized parallel implementation [ABM+ar]. At present, we are extending and improving these approaches with respect to both basic methodology and algorithm engineering and work towards a production-quality software implementation (see `https://bitbucket.org/whatshap/whatshap`).

3.4 Haplotype Reconstruction—Viral Quasispecies.

Viruses like HIV exhibit a fast mutation rate and hence evolve within a host. As a result, the host is not infected by a single virus type, but by a population of genetically diverse viruses, called a *viral quasispecies* [VSA+06]. Knowledge of the spectrum of present virus haplotypes and their relative

abundances can be important for the choice of treatment. On current second-generation sequencing machines, such a virus population can be sequenced to very deep coverage at moderate cost. Reconstructing haplotypes from the resulting sequencing reads is computationally challenging, see [BGRM12]. In prior work, we met these challenges and introduced a haplotype reconstruction algorithm that is able to reconstruct full-length haplotypes and to deliver error rates that are are lower by about two orders of magnitude compared to previous approaches on simulated data [TMB+14]. In a current project, we apply algorithm engineering techniques to speed-up the enumeration of maximal cliques, which is the core algorithmic component of this method. Moreover, we study and address artifacts present in real sequencing data and reconstruct the quasispecies of a large cohort of patient plasma samples provided by our collaborators.

3.5 Computational Pan-Genomics.

Many bioinformatics methods use the reference genome of a species under study. The used reference genomes are linear, i.e. they consist of one DNA sequence per chromosome. For instance, programs to align next-generation sequencing reads will map the reads to such a linear reference genome. Likewise, tools to call variants like SNPs and structural variations do that with respect to this reference genome. Today, however, information on common and rare variants is available for many species (and, most prominently, for *Homo sapiens*). To leverage this additional information, linear reference genomes should be replaced by variant-aware reference genomes, which comes with considerable computational challenges. We develop data structures and algorithms to overcome these challenges.

Together with four co-applicants (Victor Guryev, Alexander Schönhuth, Fabio Vandin, and Kai Ye), we successfully applied at the Lorentz Center (Leiden, Netherlands) to host a workshop on this topic. The workshop was held in June 2015 and enjoyed the participation of many internationally renowned scientists[1]. At this very productive meeting, the participants drafted a white paper summarizing the state-of-the-art and pointing out future challenges in computation pan-genomics, to be submitted soon.

References

[ABM+ar] Marco Aldinucci, Andrea Bracciali, Tobias Marschall, Murray Patterson, Nadia Pisanti, and Massimo Torquati. High-Performance Haplotype Assembly. In *Proceedings of the 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB)*, LNBI, Cambridge, UK, June to appear. Springer.

[ACE11] Can Alkan, Bradley P. Coe, and Evan E. Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, May 2011.

[BGRM12] Niko Beerenwinkel, Huldrych F. Gnthard, Volker Roth, and Karin J. Metzner. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology*, 3, September 2012.

[GCR14] Gustavo Glusman, Hannah C. Cox, and Jared C. Roach. Whole-genome haplotyping approaches and genomic medicine. *Genome Medicine*, 6(9):73, September 2014.

[KFH+15] Wigard P. Kloosterman, Laurent C. Francioli, Fereydoun Hormozdiari, Tobias Marschall, Jayne Y. Hehir-Kwa, Abdel Abdellaoui, Eric-Wubbo Lameijer, Matthijs H. Moed, Vyacheslav Koval, Ivo Renkens, Markus J. van Roosmalen, Pascal Arp, Lennart C. Karssen, Bradley P. Coe, Robert E. Handsaker, Eka D. Suchiman, Edwin Cuppen, Djie T. Thung, Mitch McVey, Michael C. Wendl, Genome of the Netherlands Consortium, Andre Uitterlinden, Cornelia M. van Duijn, Morris Swertz, Cisca Wijmenga, Gertjan van Ommen, Eline Slagboom, Dorret I. Boomsma, Alexander Schönhuth, Evan E. Eichler, Paul I. W. de Bakker, Kai Ye, and Victor Guryev. Origin, frequency and functional impact of de novo structural changes in the human genome. *Genome Research*, 25:792–801, mar 2015.

[1] See `http://www.lorentzcenter.nl/lc/web/2015/698/participants.php3?wsid=698&venue=Oort`

[KUA+07] Jan O. Korbel, Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim, Dean Palejev, Nicholas J. Carriero, Lei Du, Bruce E. Taillon, Zhoutao Chen, Andrea Tanzer, A. C. Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P. Carter, Matthew E. Hurles, Sherman M. Weissman, Timothy T. Harkins, Mark B. Gerstein, Michael Egholm, and Michael Snyder. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science*, 318(5849):420–426, October 2007.

[LMP+15] Wai Y. Leung, Tobias Marschall, Yogesh Paudel, Laurent Falquet, Hailiang Mei, Alexander Schönhuth, and Tiffanie Y. Maoz. SV-AUTOPILOT: optimized, automated construction of structural variation discovery and benchmarking pipelines. *BMC Genomics*, 16(1):238, March 2015.

[MCC+12] Tobias Marschall, Ivan G. Costa, Stefan Canzar, Markus Bauer, Gunnar W. Klau, Alexander Schliep, and Alexander Schönhuth. CLEVER: clique-enumerating variant finder. *Bioinformatics*, 28(22):2875–2882, November 2012.

[MHS13] Tobias Marschall, Iman Hajirasouliha, and Alexander Schönhuth. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics*, 29(24):3143–3150, dec 2013.

[MSB09] Paul Medvedev, Monica Stanciu, and Michael Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nat Meth*, 6(11s):S13–S20, November 2009.

[MWS+11] Ryan E. Mills, Klaudia Walter, Chip Stewart, Robert E. Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R. Keira Cheetham, Asif Chinwalla, Donald F. Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M. Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M. Kidd, Miriam K. Konkel, Joshua Korn, Ekta Khurana, Deniz Kural, Hugo Y. K. Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng Jasmine Mu, James Nemesh, Heather E. Peckham, Tobias Rausch, Aylwyn Scally, Xinghua Shi, Michael P. Stromberg, Adrian M. Stutz, Alexander Eckehart Urban, Jerilyn A. Walker, Jiantao Wu, Yujun Zhang, Zhengdong D. Zhang, Mark A. Batzer, Li Ding, Gabor T. Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E. Eichler, Mark B. Gerstein, Matthew E. Hurles, Charles Lee, Steven A. McCarroll, and Jan O. Korbel. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, February 2011.

[PMP+15] Murray Patterson*, Tobias Marschall*, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W. Klau, and Alexander Schönhuth. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Proceedings of RECOMB / Journal of Computational Biology*, 22(6):498–509, feb 2015.

[The14] The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46:818–825, June 2014.

[TMB+14] Armin Töpfer, Tobias Marschall, Rowena A. Bull, Fabio Luciani, Alexander Schönhuth, and Niko Beerenwinkel. Viral Quasispecies Assembly via Maximal Clique Enumeration. *Proceedings of RECOMB / PLoS Computational Biology*, 10(3):e1003515, mar 2014.

[VSA+06] Marco Vignuzzi, Jeffrey K. Stone, Jamie J. Arnold, Craig E. Cameron, and Raul Andino. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, 439(7074):344–348, January 2006.

[WMSM15] Roland Wittler*, Tobias Marschall*, Alex Schönhuth, and Veli Mäkinen. Repeat- and Error-Aware Comparison of Deletions. *Bioinformatics*, advance online, 2015.

PeerJ PrePrints

Statistical models of non-coding RNA-mediated gene regulation

Annalisa Marsico
Max Planck Institute for Molecular Genetics, Free University of Berlin
marsico@molgen.mpg.de

PeerJ PrePrints

1 Abstract

In our group, called RNA Bioinformatics, we are interested in investigating regulation and function of non-coding RNA transcripts by means of *in silico* methods. In particular, we are interested in those non-coding RNAs which act as regulators of gene expression, and their interplay with Transcription Factors (TFs), epigenetic marks and RNA Binding Proteins (RBPs). High-throughput experiments provide a rich source of information: the integration and interpretation of different genomic data requires the development of adequate statistical models and algorithms to uncover the putative role of non-coding transcripts in regulatory network. Regularized or sparse regression models are the main methods we employ in order to derive mechanistic hypothesis about non-coding RNA function, which can be later tested in the wet lab. We further apply unsupervised methods, such as k-means clustering or spectral clustering to characterize the properties of new non-coding RNA sub-classes by integrating several sources of high-throughput genomic data.

2 Introduction

In the cell, genomic DNA is transcribed into various types of RNA, but not all RNAs are translated into proteins. Over the past few years it has been observed, thanks to high-throughput sequencing methods, that a big portion of the human genome is transcribed in a tissue- and time-specic manner [ea13c]. Most the detected transcripts in mammals and other complex organisms are non-coding RNAs (ncRNAs), RNAs that do not encode for proteins [Mat06]. Although the functional consequences of different ncRNA classes are not yet fully understood, this does not mean that they do not contain information nor have functions.

Among the different classes of non-coding transcripts, microRNAs (miRNAs), small RNAs of 18 to 24 nucleotides in length, that post-transcriptionally regulated gene expression, are the most widely studied, but other classes of small non-coding RNAs have been characterized, such as snoRNAs (many of which still remain to be identified), snRNAs and piRNAs [ea12b]. At post-transcriptional level, about half of the human genes are regulated by microRNAs (miRNAs), which can bind to the 3-untranslated regions (3' UTRs) or coding regions of target genes, leading to the degradation of target mRNAs or translational repression [ea14a]. MiRNAs are associated with an array of biological processes, such as embryonic development and stem cell functions in mammals [ea09], and a crucial role of miRNAs in gene regulatory networks has been recognized in the last decade in the context of cancer development, as well as immune system response [ea06]. Given the growing importance of miRNA function in contributing to the control of gene expression, most of the research in the past decade has been focusing on miRNA-gene target prediction and gene regulatory networks have been expanded to include the involvement of miRNAs.

Despite great progress in understanding the biological role of miRNAs, our understanding of how miRNAs are regulated and processed is still developing. High-throughput sequencing data have provided a robust platform for transcriptome-level, as well as gene-promoter analyses. Some recent *in silico* predictive models for miRNA promoter recognition enable the challenging task of locating the Transcription Start Sites (TSSs) of transient miRNA primary transcripts, thereby allowing the prediction of their

regulatory elements and Transcription Factor Binding Sites (TFBSs) [ea11a, ea13a, ea14b].

Besides small non-coding RNAs, advances in high-throughput sequencing, combined with genome-wide mapping of chromatin modification signatures and bioinformatics pipelines, have resulted in the identification of tens of thousands of longer non-coding transcripts (including intergenic and overlapping sense and antisense transcripts), whose functional significance is still controversial [ea15]. Although the existence of such non-coding RNAs has been validated in multiple experimental systems, and several long intergenic non-coding RNAs (lincRNAs) have been described in processes of gene silencing [ea10b], imprinting [ea10c], and lately in gene activation [ea10d, ea13b], a global spectrum of all possible lincRNA-related functions, as well as automated methods for lincRNA function prediction are missing. This is mainly due to the fact that the vast majority of lincRNAs shows little evidence of evolutionary conservation at sequence level [ea11b]. The development of systematic statistical methods to find significant associations or co-expression between protein-coding genes and lincRNAs, as well as the inspection of evolutionary signatures based on structure rather than sequence motifs is needed to help inferring many still-unknown lincRNA functions.

PeerJ PrePrints

3 Research concept

In our group we are interested in the mechanisms that govern the regulation of non-coding RNAs, as well as functional analysis of different classes of non-coding RNAs, with focus on miRNAs and lincRNAs. To these goals, we develop statistical models and use existing machine learning approaches to answer questions like: How can miRNA promoters be detected genome-wide and distinguished from transcriptional noise? What are the genomic features that control miRNA processing? What are the genes regulated by a certain long non-coding RNA? How can we automatically classify long non-coding RNAs into functional classes based on sequence / structure features? Part of the group focuses on statistical modeling of miRNA regulatory elements from next-generation sequencing data, such as expression data, epigenetics marks and genetic variants that control miRNA expression. The other part of the group focuses on the characterization of lincRNA function using methods, such as sparse regression, network analysis and data integration.

3.1 Statistical modeling of miRNA Biogenesis

MiRNAs are regulated at different levels during their biogenesis pathway [Mat06]. Understanding which TFs regulate a certain miRNA at a certain time in a certain tissue requires the knowledge of the location of the core promoter of the miRNA primary transcript. In my previous work, I developed a semi-supervised machine learning method for miRNA promoter recognition called PROmiRNA [ea13a]. The PROmiRNA mixture model assumes that the read count distribution observed from deepCAGE data is represented by a mixture of putative promoters versus background noise. In order to not underestimate the number of true promoters (due to lowly expressed transcripts), sequence features, such as CpG content, conservation and TATA box affinity are introduced into the model through an informative prior. During training, known miRNA promoters are included as exact examples and the output of the classifier is a posterior probability for a certain regions to be a real promoter. The application of PROmiRNA to the human genome allowed us for the first time to study the characteristics of regulatory elements of different miRNA promoter classes. We are currently working on a more scalable version of the PROmiRNA software, as well as a web-server for allowing the exploration of miRNA promoters across different tissues. By exploiting the functionalities of PROmiRNA we are currently able to explore other aspects of miRNA regulation. Ongoing projects in the lab include 1) the development of a regression model (elastic net) for predicting miRNA expression based on chromatin signatures around at the predicted promoter regions and 2) the prediction of causal genetic variants (eQTL SNPs) which alter miRNA expression in promoter regions by using different regulatory elements as covariates.

Preliminary results indicate a crucial role of DNA methylation in shaping miRNA expression and provide a list of causal genetic variants localized in proximity of tissue-specific miRNA promoters.

Global mature miRNA expression is not only regulated at transcriptional level, but several post-transcriptional steps influence the final miRNA expression level. In our previous work, together with our experimental partners from the group of Dr. Ulf Orom, we have defined a quantitative measure of miRNA processing from RNA-Seq data and built a classification model to discriminate efficient from non-efficient processing based on sequence features, i.e. specific and degenerate k-mers [ea14c]. Prompted by the results of this study we are currently investigating the relationship between miRNA processing and epigenetic signatures of miRNA genes.

3.2 Function prediction of long non-coding RNAs

Recent studies have reported enhancer functions of long non-coding RNAs [ea10d], pointing to active transcription of previously identified enhancer regions. In order to identify specific signatures in this new class of enhancer lincRNAs, in one of our current projects we have collected transcriptome data and epigenetics marks in MCF7 cells and identified, by means of k-means clustering, about 400 lincRNAs with putative active enhancer function. We could also associate this cluster of lincRNAs to high hypomethylation specificity among cell lines and to high co-variation in the expression of their nearby genes, supporting further their role as putative enhancers activated by a methylation-dependent mechanism. Motivated by the fact that if the expression of a gene and a long non-coding RNA co-vary among several tissues, then a direct or indirect association can be inferred between the two, we try to infer putative top ranking associations between genes and long non-coding RNAs across different tissues, Given that the number of variables (all annotated genes and lincRNAs) is much higher than the number of samples (different tissues with available expression data) we use sparse regression techniques, such as Orthogonal Matching Pursuit to prioritize significant interactions.

Long ncRNAs have been shown to physically connect the genomic regions of regulated genes with their own genomic locus, thereby mediating gene activation or enhancer function through direct chromatin interactions [ea10d]. Such data are useful to detect direct interactions, therefore we are currently integrating freely available chromatin-conformation data, such as ChIA-PET data [ea10a], with chromatin states to build the physical interaction network involving genes, long non-coding RNAs and other putative regulatory elements in a specific tissue. Such retrieved interactions can be converted to a weighted adjacency matrix, which we then analyze by means of Spectral Clustering in order to identify potentially important regulatory modules which involve lincRNAs.

Long non-coding RNAs do not act alone to perform their activating or repressing function but often associate with RNA-binding proteins or chromatin remodeling complexes that guide them to their sites of action. Identifying which proteins bind to a specific long non-coding RNA can help shedding light on its function. Interactions of long non-coding RNAs with RNA-binding proteins can be detected via technologies such as CLIP-seq [ea12a]. The technology is really new and so far few methods have been developed to reliably identify binding sites above noise and in the presence of appropriate control, in particular for iCLIP data [ea12a]. Although this project just started, our idea is to model the read count distribution for a certain experiment and the control simultaneously by means of a factorial Hidden Markov Model, taking into account special features of iCLIP experiments (e.g. truncation rates, sequence bias) as additional covariates.

4 Outlook

In summary, our group is working towards a global understanding of how non-coding RNAs, such as miRNAs and long non-coding RNAs, participate in gene regulatory networks. We employ several

PeerJ PrePrints

machine learning methods, such as semi-supervised or supervised classification models to characterize promoters and regulatory features of miRNAs. Together with the group of Bernd Schmeck at the Uniklinukum Marburg (SFB TR84), we will apply our model in the context of infectious diseases, to elucidate the regulatory mechanisms of miRNAs induced in the host cells by a specific infectious process. Our analysis be will extended to include the possible role of lincRNAs in shaping the regulatory network activated by the host in response to the pathogenic infection, with the hope to discover new functions for long non-coding RNAs. In order to unravel the mechanisms of long non-coding RNA function we would like to discover structural motifs, as well as common RNA-binding protein sites among long non-coding RNAs with similar expression/activation patterns. If we can find signatures or structure/sequence motifs among 'related' lncRNAs, these patterns could hint to the lncRNA function and would represent a first step towards a systemathic functional classification of long non-coding RNAs.

References

[ea06] A. Esquela-Kerscher et al. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer*, 6:259–269, 2006.

[ea09] D.P. Bartel et al. MicroRNAs: target recognition and regulatory functions. *Cell*, 136:215–233, 2009.

[ea10a] G. Li et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genom Biol*, 11(1):R22, 2010.

[ea10b] M. Huarte et al. A large intergenic non-coding RNA induced by p53 mediates global gene expression in p53 response. *Cell*, 142(3):409–419, 2010.

[ea10c] M. Huarte et al. Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–1076, 2010.

[ea10d] U.A. Orom et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143(1):46–58, 2010.

[ea11a] C. Chien et al. Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucl Acids Res*, 39(21):9345–56, 2011.

[ea11b] I. Ulitsky et al. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147(7):1537–50, 2011.

[ea12a] J. Koenig et al. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet*, 13(2):77–83, 2012.

[ea12b] M.S. Kowalczyk et al. Molecular biology: RNA discrimination. *Nature*, 482:310–311, 2012.

[ea13a] A. Marsico et al. PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol*, 14:R84, 2013.

[ea13b] E.G. Berghoff et al. Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. *Development*, 140:4407–4416, 2013.

[ea13c] I. Ulitsky et al. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*, 154(1):26–46, 2013.

[ea14a] B.N. Davis et al. Regulation of MicroRNA Biogenesis: A miRiad of mechanisms. *Commun Signal*, 10(7):7–18, 2014.

[ea14b] G. Georgakilas et al. microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat Commun*, page doi:10.1038/ncomms6700, 2014.

[ea14c] T. Conrad et al. Microprocessor activity controls differential miRNA biogenesis In Vivo. *Cell Rep*, 9:542–554, 2014.

[ea15] J.S. Mattick et al. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol*, 22(1):5–7, 2015.

[Mat06] J.S. Mattick. Non-coding RNA. *Hum Mol Genet*, 15(Supp1):R17–R29, 2006.

PeerJ PrePrints

Statistical Learning in Computational Biology

Nico Pfeifer
Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics
npfeifer@mpi-inf.mpg.de

Group Development

Nico Pfeifer is at the MPI for Informatics since October 2011. He started his group in January 2013 and became a senior researcher in November 2014 having the right to grant Ph.D. titles from Saarland university. Nico Pfeifer supervises five Ph.D. students directly and co-supervises one Ph.D. student together with Thomas Lengauer.

Vision and Research Strategy

Recent advances in high-throughput technologies have led to an exponential increase in biological data (such as genomic, epigenomic and proteomic data). To gain meaningful insights in such large data collections, efficient statistical learning methods are needed that take into account various sources of confounding such as batch effects or population structure, inherent to large biological data sets. We are interested in developing and applying new machine learning / statistical learning methods to solving computational biology problems and answering new biological questions. Application areas include the study of viruses like HIV, Hepatitis C or Influenza as well as the field of epigenetics. Method-wise we are interested in

- integration of heterogeneous data sets
- improving interpretability of non-linear estimators
- efficient learning methods for large data sets.

Due to about two million new HIV infections per year and about 35 million people living with an HIV infection world-wide, the HI virus is still a major threat to mankind. Two areas are of particular importance:

- research towards a vaccine against HIV
- personalized HIV treatment

We are conducting research in both of these areas. Examples include modeling the adaptation of HIV in response to external pressure by the immune system [YKK$^+$13, CBM$^+$12, CLP$^+$12], building better and more interpretable predictors for HIV coreceptor usage and CCR5 antagonist resistance prediction [PL12] as well as the analysis of potent broadly HIV-1 neutralizing antibodies ([PWL14]).
We investigated certain biases in biological data that may have very important implications for the interpretation of results based on this data (see [PWL14] and [DGG$^+$15]).
Additionally, we are also working on methods that can better deal with noisy data. One application scenario is the analysis of molecular measurements on cancer samples. Here, many effects can introduce biases (e.g., population structure, cryptic relatedness, batch effects). If one builds a prediction tool with the assumption that new data to come will be very similar to the data on which the model is trained, standard approaches are applicable. Unfortunately, this is not very often the case. Therefore, we introduced a method that is able to estimate certain differences in the underlying distribution of the training data and the test data and correct for them in the final prediction method. Furthermore, we provided interpretable results that can be used to understand the underlying causes of the prediction label (see [JP14]).

PeerJ PrePrints

Another important area is how to best integrate the different measurements (e.g., gene expression, DNA methylation, copy number variation). Here we extended methods for unsupervised multi kernel learning to deal with the different data types ([SP15]).
Additionally, we are interested in developing methods for the analysis of open chromatin regions as well as the three-dimensional organization of chromosomes.

Selected Research Projects

Statistical Learning for Visualizing, Analyzing and Integrating Different Omics Data Sets

Over the past decades, biology has transformed into a high throughput research field both in terms of the number of different measurement techniques as well as the amount of variables measured by each technique (e.g., from Sanger sequencing to deep sequencing) and is more and more targeted to individual cells [SBL13]. This has led to an unprecedented growth of biological information. Consequently, techniques that can help researchers find the important insights of the data are becoming more and more important. Molecular measurements from cancer patients such as gene expression and DNA methylation are usually very noisy. Furthermore, cancer types can be very heterogeneous. Therefore, one of the main assumptions for machine learning, that the underlying unknown distribution is the same for all samples in training and test data, might not be completely fulfilled.

Interpretable per Case Weighted Ensemble Method for Cancer Associations

We introduced a method that is aware of the potential bias regarding different batches of data and utilizes an estimate of the differences during the generation of the final prediction model. For this, we introduced a set of sparse classifiers based on L1-SVMs [BM98], under the constraint of disjoint features used by classifiers. Furthermore, for each feature chosen by one of the classifiers, we introduced a regression model based on Gaussian process regression that uses additional features. For a given test sample we can then use these regression models to estimate for each classifier how well its features are predictable by the corresponding Gaussian process regression model. This information is then used for a confidence-based weighting of the classifiers for the test sample. Schapire and Singer showed that incorporating confidences of classifiers can improve the performance of an ensemble method [SS99]. However, in their setting confidences of classifiers are estimated using the training data and are thus fixed for all test samples, whereas in our setting we estimate confidences of individual classifiers per given test sample.

In our evaluation, the new method achieved state-of-the-art performance on many different cancer data sets with measured DNA methylation or gene expression. Moreover, we developed a method to visualize our learned classifiers to find interesting associations with the target label. Applied to a leukemia data set we found several ribosomal proteins associated with leukemia that might be interesting targets for follow-up studies and support the hypothesis that the ribosomes are a new frontier in gene regulation. This research project was presented at WABI 2014 [JP14].

Integrating Different Data Types by Regularized Unsupervised Multiple Kernel Learning with Application to Cancer Subtype Discovery

Despite ongoing research, cancer remains a major health threat. The identification of subtypes of tumors in certain tissues can guide the decision which treatment may be beneficial for the respective patient. Nowadays established cancer subtypes are mainly based on individual types of molecular data, such as gene expression or DNA methylation. However, the analysis of multidimensional data, consisting of measurements using different platforms, may reveal intrinsic characteristics of the tumor which are

PeerJ PrePrints

based on dependencies between these different data types and can therefore only be detected when integrating the available information. Large-scale projects, such as The Cancer Genome Atlas (TCGA) [TCG] accumulate such heterogeneous data for various cancer types, but we still lack computational methods that are able to reliably integrate the given data.

To enable integrative, exploratory data analysis, we extended an approach to unsupervised multiple kernel learning for dimensionality reduction [LLF11]. In a first step, each input data type is represented by one or several kernel matrices. At this point, a major advantage is the ability of the method to automatically weight the kernel matrices, such that the user is alleviated from the burden of deciding on a kernel function or kernel parameters for each data type, instead, one can simply input a set of kernel matrices for each data type and let the method determine the optimal weighting. In an iterative optimization process, the method then trains a kernel weight vector β, used to calculate the weighted linear combination of the input kernels, and a projection matrix A which allows for dimensionality reduction. Due to the graph embedding framework [YXZ$^+$07] which forms the basis of the method, a large number of dimensionality reduction methods can be applied.

We applied this method to patient data of five different cancer types (glioblastoma multiforme, breast invasive carcinoma, kidney renal clear cell carcinoma, lung squamous cell carcinoma, and colon adenocarcinoma), where for each cancer type three different data types (gene expression, DNA methylation, and miRNA expression) were available. For dimensionality reduction we applied the locality preserving projections algorithm [HN04], which is based on the k-nearest neighborhood of a sample. We used radial basis kernel functions and, in order to investigate the efficacy of the kernel weighting, we compared two different scenarios. In the first one, we represent each input type as one kernel matrix. In Scenario 2, we use five different kernel matrices per data type, obtained by using five different kernel parameters. Our analysis revealed, that uninformative input kernel matrices indeed hardly influence the ensemble matrix. Subsequently, we applied k-means clustering to the integrated patient data to identify integrated cancer subtypes. In order to assess the biological validity of these clusters, we performed a survival analysis evaluating if the potential subtypes differ in prognosis. In Scenario 1 (one kernel per data type), we found significant differences in survival time between the subtypes for all but one cancer type. With Scenario 2, the significance for most data sets increased such that the identified subtypes are at least as significant as those identified by state-of-the-art methods, i.e., the clusters obtained reflect a better separation according to survival time of the patients than the results obtained in Scenario 1. Moreover, a leave-one-out cross validation approach showed, that the identified subtypes are relatively stable, with no decrease in stability when using more than one kernel matrix for a data type. We further looked into the groups identified for glioblastoma multiforme. For this cancer type, we were able to find subtypes that are established for distinct individual data types, but also additional subtypes, potentially based on interaction of the integrated data types. For glioblastoma multiforme, we also investigated how the subtypes respond to different treatments. For the drug Temozolomide, patients from certain subtypes seemed to benefit from that therapy, appearing a significantly increased survival time compared to patients from the same subtype but not treated with Temozolomide. In other clusters, no significant survival time differences between patient treated and not treated with this drug were observed. Overall, our method shows promising results when applied in the field of cancer subtype identification.
A manuscript describing the work was presented at ISMB/ECCB 2015 [SP15].

Projects and Cooperations

We are collaborating with several researchers internationally, nationally and also on campus: David Heckerman, Microsoft Research, Jonathan Carlson, Microsoft Research, Anne-Mieke Vandamme, KU Leuven, Rolf Kaiser, University of Cologne, Jörn Walter, University of the Saarland, Marcel Schulz, MMCI, Saarbrücken, Olga Kalinina, MPI for Informatics

PeerJ PrePrints

References

[BM98] Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

[CBM+12] Jonathan M Carlson, Chanson J Brumme, Eric Martin, Jennifer Listgarten, Mark A Brockman, Anh Q Le, Celia K S Chui, Laura A Cotton, David J H F Knapp, Sharon A Riddler, Richard Haubrich, George Nelson, Nico Pfeifer, Charles E Deziel, David Heckerman, Richard Apps, Mary Carrington, Simon Mallal, P Richard Harrigan, Mina John, and Zabrina L Brumme. Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. *Journal of virology*, 86(24):13202–16, December 2012.

[CLP+12] Jonathan M Carlson, Jennifer Listgarten, Nico Pfeifer, Vincent Tan, Carl Kadie, Bruce D Walker, Thumbi Ndung'u, Roger Shapiro, John Frater, Zabrina L Brumme, Philip J R Goulder, and David Heckerman. Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *Journal of virology*, 86(9):5230–43, May 2012.

[DGG+15] Matthias Döring, Gilles Gasparoni, Jasmin Gries, Karl Nordström, Pavlo Lutsik, Jörn Walter, and Nico Pfeifer. Identification and Analysis of Methylation Call Differences Between Bisulfite Microarray and Bisulfite Sequencing Data with Statistical Learning Techniques. *BMC Bioinformatics (Proc. ISCB)*, 16(Suppl 3), 2015.

[HN04] Xiaofei He and Partha Niyogi. Locality Preserving Projections. In S. Thrun, L.K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 153–160. MIT Press, 2004.

[JP14] Adrin Jalali and Nico Pfeifer. Interpretable Per Case Weighted Ensemble Method for Cancer Associations. In Dan Brown and Burkhard Morgenstern, editors, *Algorithms in Bioinformatics (WABI 2014)*, volume 8701 of *Lecture Notes in Bioinformatics*, pages 352–353, Wroclaw, Poland, 2014. Springer.

[LLF11] Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh. Multiple Kernel Learning for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1147–1160, 2011.

[PL12] Nico Pfeifer and Thomas Lengauer. Improving HIV Coreceptor Usage Prediction in the Clinic Using hints from Next-generation Sequencing Data. *Bioinformatics*, 28(18):i589–i595, 2012.

[PWL14] Nico Pfeifer, Hauke Walter, and Thomas Lengauer. Association Between HIV-1 Coreceptor Usage and Resistance to Broadly Neutralizing Antibodies. *Journal of Acquired Immune Deficiency Syndromes : JAIDS*, 67(2):107–112, 2014.

[SBL13] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet.*, 14(9):618–30, September 2013.

[SP15] Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics (Oxford, England)*, 31(12):i268–i275, June 2015.

[SS99] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[TCG] The Cancer Genome Atlas, Website, Available from: http://cancergenome.nih.gov/.

[YKK+13] Yuichi Yagita, Nozomi Kuse, Kimiko Kuroki, Hiroyuki Gatanaga, Jonathan M Carlson, Takayuki Chikata, Zabrina L Brumme, Hayato Murakoshi, Tomohiro Akahoshi, Nico Pfeifer, Simon Mallal, Mina John, Toyoyuki Ose, Haruki Matsubara, Ryo Kanda, Yuko Fukunaga, Kazutaka Honda, Yuka Kawashima, Yasuo Ariumi, Shinichi Oka, Katsumi Maenaka, and Masafumi Takiguchi. Distinct HIV-1 Escape Patterns Selected by Cytotoxic T Cells with Identical Epitope Specificity. *Journal of virology*, 87(4):2253–63, February 2013.

[YXZ+07] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

PeerJ PrePrints

Management of simulation studies in computational biology

Dagmar Waltemath
e:Bio Junior Research Group SEMS
Systems Biology and Bioinformatics, Institute of Computer Science, University of Rostock, Germany
dagmar.waltemath@uni-rostock.de

PeerJ PrePrints

1 On the need for data management in computational biology projects

Data management is a well defined task in computer science which investigates methods for organising and controlling the information generated during (research) projects. It comprises several tasks, including data storage, search, retrieval, version control and provenance. Effective data management strategies for computational biology are needed to handle the increasing amount of data that is being generated and processed: High-throughput experiments generate large amounts of data; computational models become complex; novel methods for model coupling enable researchers to combine models into even larger systems; increasing computational power allows for complex simulations; and the availability of data at different scales demands clever integration techniques. However, recent studies showed that the rate of reproducibility of scientific results in the life sciences, including computational biology, is not acceptable [Ioa14]. As a consequence, efforts have been launched to improve reusability and reproducibility of biomedical results (e. g., [M$^+$14, Ioa14]), and results of simulation studies in particular [W$^+$11a, B$^+$14, C$^+$15]. Today, paths towards improved data management are discussed by funders and publishers, in large scale projects and by individual researchers. For example, funders established policies such as the *ERASysAPP Data Management Guidelines*; the German Network for Bioinformatics Infrastructure, de.NBI (http://www.denbi.de) has dedicated data management centers; and projects are funded to develop support for sustainable data management, e. g., FAIR-DOM (http://fair-dom.org).

The junior research group SEMS (http://sems.uni-rostock.de) focuses on the management of specific data: It develops methods and tools for the management of simulation studies in computational biology.

2 Methods and tools for the management of simulation studies

SEMS focuses on models encoded in XML standard formats and annotated with terms from bio-ontologies. More specifically, we work with models encoded in the *Systems Biology Markup Language* (SBML [H$^+$03]) and CellML [C$^+$03]. The majority of these models are mathematical models describing biological and physiological processes. The execution of these models can be described using the *Simulation Experiment Description Markup Language* (SED-ML [W$^+$11b]). While these three formats encode for the necessary information to run models [W$^+$11a], additional semantic annotations are needed to capture the biology [C$^+$11b], for example annotations to the Gene Ontology. Graphical representations of the networks can be standardised using the Systems Biology Graphical Notation (SBGN [LN$^+$09]). Together with ongoing developments of standards for data representation and ontologies to express the behavior and dynamics of a model, a whole plethora of data is collected when performing a simulation study. Figure 1 summarises how SEMS supports the management and integration of that data: The displayed reaction is part of a model reproducing the mitotic oscillator involving Cyclin and cfc2 kinase. The model itself was published in an article by Goldbeter in 1991 [Gol91]. Its SBML encoding, together with the graphical network in SBGN, is provided through BioModels Database, a rich resource of curated and annotated models [L$^+$10]. A standardised drawing of the interactions in the network enables researchers to quickly grasp the essence of what the model encodes. It is particularly useful to discuss different versions of the model with collaborators in large projects. Models may be simulated in different ways, with the actual setup of the experiments depending on the specific question asked.

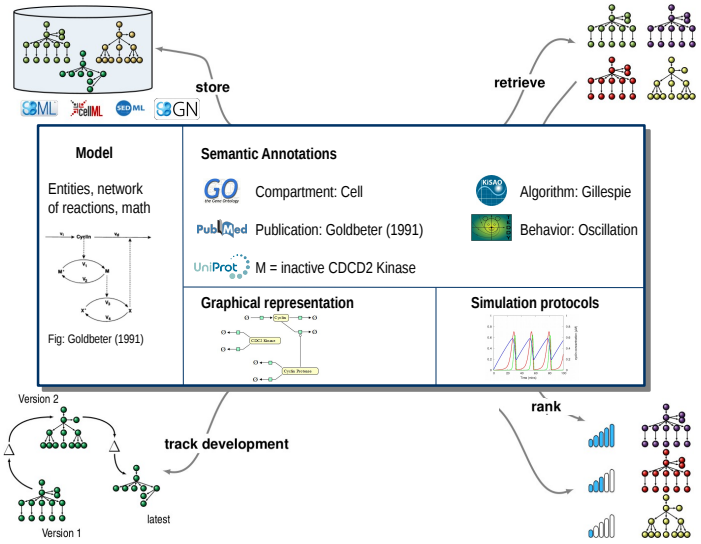


Figure 1: **Overview of integrated, model-related data and developed model management solutions.** The inner box shows the different types of model-related data that we integrate on the storage layer: model code (SBML, CellML), reference publications (e. g., PubMed), semantic annotations (bio-ontologies), graphical representations (SBGN), and simulation protocols (SED-ML). The outer ring shows our contributions to four major model management tasks: storage, search and retrieval, ranking, and version control.

These variations can be stored in SED-ML files, together with information on the simulation algorithm to use, or with links to parametrisation files and result data. An ontology of simulation algorithms is the *Kinetic Simulation Algorithm Ontology* (KiSAO [C$^+$11b]). Each result can be linked to a defined behavior encoded in the *Terminology for the Description of Dynamics* (TEDDY, [C$^+$11b]). Finally, the reference publication, typically in PDF format, is linked through a document object identifier (DOI).

The junior research group SEMS works towards better reproducibility of simulation studies, to lower the effort of reusing existing work, and to make scientific investigations more open and transparent. As depicted in Figure 1, we apply methods from data management on the problem of model management in computational biology. We are currently supported by three BMBF grants: The e:Bio junior group SEMS itself, one project in the German infrastructure for Bioinformatics (de.NBI), and another e:Bio project on SBGN-ED, an editor for SBGN maps. In all three projects, we collaborate closely with developers of model repositories (BioModels Database, *Physiome Model Repository* (PMR2 [Y$^+$11])) and data management systems such as SEEK [W$^+$15]. With our methods and tools, we aim to easen the findability, comparison, exploration, and understanding of simulation studies in computational biology.

Model search and retrieval The large number of published models necessitates smart search engines that incorporate the context of the model, semantic annotations, structural information and even allow for sub-model search. The results of a search need to be ranked according to user preferences. In 2010, we introduced the *ranked retrieval* engine for models, MORRE [H$^+$10]. It is a method to search and retrieve models from a given set, and to rank the results using state-of-the-art Information Retrieval methods. We showed how such a system improves the search in BioModels Database and PMR2. The first version of MORRE already considered model encoding and semantic annotations. We recently extended the method to enable clustering of models based on annotations [A$^+$15]. Furthermore, we investigate how similarity can be determined by transforming the network structure into a bipartite graph and detecting subgraph isomorphisms [RW14].

Model version control and provenance Models evolve over time, e. g., during model design, model publication, curation and reuse. A version of a model may fit a purpose while another may not, and

PeerJ PrePrints

a model update may lead to modifications in the obtained simulation results. Model version control is therefore a necessary feature of all tools offerering model code for reuse. It allows users to track and understand the changes in a model [W+13]. To this end, we developed an algorithm for difference detection in versions of models, BiVeS [S+15]. It takes two versions of an SBML- or CellML-encoded model and calculates the differences. These differences can be exported in human-readable format, as an XML diff file, or they can be displayed visually. Since PMR2 integrated BiVeS, users can explore the changes of model code between different exposures. In the functional curation framework [C+11a], the BiVeS webservice is used to display differences in versions of CellML models.

Integration of model-related data We investigated the use of graph databases for the management of model-related data files. Our prototype system, MASYMOS [H+15], exemplifies how model-related data can be stored and linked, thereby enabling the retrieval of complete simulation studies. As most data are already encoded in XML, they can easily be converted into graph-like representations. Graph databases, in addition, enable flexible linking of data items. MASYMOS can thus reflect facts such as that a model is linked to several experiments, or that particular model entities are observed in a simulation. Ultimately, the application of graph concepts enables novel types of queries, for example for sub-models. This again can have a positive effect on the results of the ranked retrieval.

Exchange of reproducible simulation studies MASYMOS and MORRE contribute to the storage and retrieval of model-related data. How can one now export the extracted studies efficiently, without loosing important files, nor extracting files in wrong versions? Over the past years we contributed to the development of the COMBINE archive [B+14]. It serves as a container for all files necessary to reproduce a simulation study. Using the archive, simulation studies can thus be shipped as one single file. We developed a set of tools that read and modify archives; generate archives from data in MASYMOS; or enable sharing of archives online [SW15]. Another contribution of the group is the implementation of support for the COMBINE archive in the functional curation framework, where users can compare how different models handle a specific simulation task [Mir15].

Contribution to standards development SEMS actively contributes to the development of community standards, Minimum Information guidelines and ontologies through the *Computational Modeling in Biology Network* (COMBINE [W+14]). Specifically, our group members are editors of SBML, co-founders of SED-ML, active developers of the COMBINE Archive standard and coordinators of the COMBINE Network. We help with organising the annual community meetings, we support grant applications and outreach activities. For example, we teach students how to transform their modeling results into standard-compliant, reproducible simulation studies, and how to publish their data openly and sustainably. Finally, our group is part of the systems biology node for data management within the de.NBI network. Here we work towards integrating our model management tools into SEEK.

3 Summary: Promoting reproducible and open science

Reproducibility of scientific results is a major challenge in computational biology. The problem is manifold and can be addressed from different angles. In SEMS, we focus on the data management aspect: Only studies that are findable, verifiable, curated and well documented can be reproduced. A prerequisite is the availability of all necessary data and in interoperable formats. To this end we develop novel methods and tools for model management, specifically for search, retrieval, ranking, version control, and integration of model-related data. Furthermore, we are actively engaged in standards development and community efforts. Goals for the forthcoming years are to integrate SEMS tools in existing data management platforms, to raise the awareness for standards and reproducible science, and to integrate further types of data, specifically biomedical and clinical data.

PeerJ PrePrints

References

[A+15] R Alm et al. Annotation-based feature extraction from sets of SBML models. *Journal of Biomedical Semantantics*, 6(1):20, 2015.

[B+14] FT Bergmann et al. COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. *BMC Bioinf*, 15(1):369, 2014.

[C+03] A Cuellar et al. An Overview of CellML 1.1, a Biological Model Description Language. *SIMULATION*, 79(12):740–747, 2003.

[C+11a] J Cooper et al. High-throughput functional curation of cellular electrophysiology models. *Progress in Biophysics and Molecular Biology*, 107(1):11–20, 2011.

[C+11b] M Courtot et al. Controlled Vocabularies and Semantics in Systems Biology. *Molecular Systems Biology*, 7, 2011.

[C+15] J Cooper et al. A call for virtual experiments: accelerating the scientific process. *Progress in Biophysics and Molecular Biology*, 117(1):99–106, 2015.

[Gol91] Albert Goldbeter. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proceedings of the National Academy of Sciences*, 88(20):9107–9111, 1991.

[H+03] M Hucka et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *BIOINFORMATICS*, 19(4):524–531, 3 2003.

[H+10] R Henkel et al. Ranked retrieval of computational biology models. *BMC Bioinformatics*, 11(1):423, 2010.

[H+15] R Henkel et al. Combining computational models, semantic annotations and simulation experiments in a graph database. *DATABASE*, 2015:bau130, 2015.

[Ioa14] J Ioannidis. How to make more published research true. *PLoS Medicine*, 11(10), 2014.

[L+10] C Li et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4:92, Jun 2010.

[LN+09] N Le Novère et al. The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741, 8 2009.

[M+14] M Macleod et al. Biomedical research: increasing value, reducing waste. *The Lancet*, 383(9912):101–104, 2014.

[Mir15] G Mirams. Introducing the 'Cardiac Electrophysiology Web Lab'. https://mirams.wordpress.com/2014/05/09/web-lab/ (last accessed 2015-06-30), 2015.

[RW14] C Rosenke and D Waltemath. How Can Semantic Annotations Support the Identification of Network Similarities? In *Proceedings of the 2014 Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*, 2014.

[S+15] M Scharm et al. An algorithm to detect and communicate the differences in computational models describing biological systems. *BIOINFORMATICS, accepted for publication*, 2015.

[SW15] M Scharm and D Waltemath. Extracting reproducible simulation studies from model repositories using the CombineArchive Toolkit. In Norbert Ritter et al., editors, *Datenbanksysteme für Business, Technolog und Web (BTW 2015) - Workshop*, volume P-242, pages 137–44, 2015.

[W+11a] D Waltemath et al. Minimum information about a simulation experiment (MIASE). *PLoS Computational Biology*, 7(4):e1001122_1–e1001122_4, 2011.

[W+11b] D Waltemath et al. Reproducible computational biology experiments with SED-ML - the simulation experiment description markup language. *BMC Systems Biology*, 5(1):198, 2011.

[W+13] D Waltemath et al. Improving the reuse of computational models through version control. *BIOINFORMATICS*, 29(6):742–748, 2013.

[W+14] D Waltemath et al. Meeting report from the fourth meeting of the Computational Modeling in Biology Network (COMBINE). *Standards in Genomic Sciences*, 9(3), 2014.

[W+15] K Wolstencroft et al. SEEK: a systems biology data and model management platform. *BMC Systems Biology*, 9(1):33, 2015.

[Y+11] T Yu et al. The physiome model repository 2. *BIOINFORMATICS*, 27(5):743–744, 2011.

PeerJ PrePrints

PeerJ PrePrints

Highlight Abstracts

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

Varying levels of complexity in transcription factor binding motifs

Jan Grau[1] and Jens Keilwagen[2]
[1] *Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany*
[2] *Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany*
grau@informatik.uni-halle.de

PeerJ PrePrints

Introduction

Transcriptional regulation mediated by transcription factors (TFs) binding to genomic DNA is one of the fundamental regulatory steps of gene expression. Over the last years, the importance of dependencies between different positions of transcription factor binding sites (TFBSs) has been debated controversially [B$^+$09a, ZS11, M$^+$11]. Several publications argue that TF-DNA binding energies can often be captured by simple weight matrices [ZS11, W$^+$13], whereas others find that considering dependencies increases the performance of TFBS predictions [M$^+$13, MW13, K$^+$13, G$^+$13].

Here [KG15], we aim at providing new insights into the importance of dependencies in transcription factor binding sites and investigate the diverse sources of such dependencies on *in-vitro* genomic context protein binding microarray (gcPBM) data [M$^+$13] and *in-vivo* ChIP-seq data from ENCODE [ENC12]. For this purpose, we propose a new class of probabilistic models that allow for learning dependencies between binding site positions discriminatively, which we call *sparse local inhomogeneous mixture* (Slim) models. For representing dependencies graphically, we develop a new visualization technique, which we call *dependency logos*.

Sparse local inhomogeneous Markov models

Determining a probabilistic model requires the selection of features and the estimation of corresponding model parameters. Typically, feature selection is performed in discrete space (features are selected or not), while parameter estimation is performed in continuous space. For parameter estimation, discriminative learning principles have been proven superior over generative ones in many areas including motif discovery [Bai11, H$^+$11, G$^+$13], but typically demand for time-consuming numerical optimization, which makes them intractable for traditional feature selection that requires a new optimization for each (promising) feature subset.

To overcome this situation, we propose Slim models that use the alternative concept of soft feature selection. More specifically, the probability of a nucleotide at a certain position of a binding site may depend on any nucleotide observed at a preceding position. Since it is unknown beforehand, which of these putative dependencies are important, the Slim model handles this information as a hidden variable resulting in a local mixture model. During the learning process, the parameters of this mixture model are adapted, such that a single position or a small subset of preceding positions obtains a large weight, whereas the others are down-weighted, yielding a soft feature selection.

Dependency logos

We present dependency logos as a new way of visualizing dependency structures within binding sites. In contrast to sequence logos, dependency logos make dependencies between binding site positions visually perceptible. In contrast to previous approaches, dependency logos are model-free and only require a set of aligned sequences, e.g., predicted binding sites, and, optionally, associated weights as input.

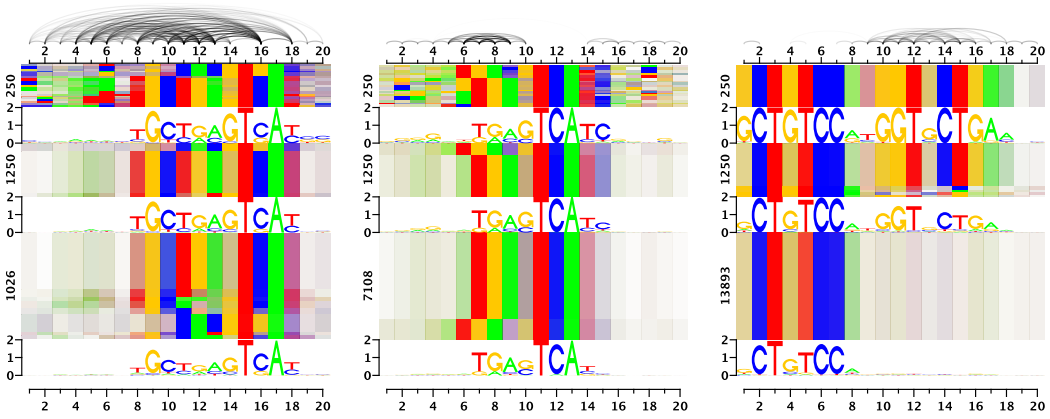
Dependency logos make dependencies between different motif positions visually perceptible by three

key ideas. First, dependency logos are directly based on binding sites instead of abstract binding motifs, e.g., mononucleotide distributions of PWM models. Second, we cluster binding sites by their nucleotides at those positions showing the strongest dependencies to other positions. If, for instance, position i shows the strongest dependencies to other positions and, of those, the dependency between position j and i is the strongest, we create at most 16 clusters according to the combinations of the two nucleotides present at positions j and i. This procedure may be repeated recursively for each of the clusters (e.g., those sequences with a TC at position j and i). Third, we visualize each cluster as one row of colored boxes using the familiar colors of sequence logos and with height proportional to cluster size. If more than one nucleotide is present at a certain binding site position in a cluster, we mix the colors representing those nucleotides and set their saturation based on information content in analogy to the height of stacks in sequence logos.

Results

We demonstrate that Slim models in combination with a discriminative learning principle yield an overall improved performance compared to state of the art tools and compared to other probabilistic models including position weight matrix models on gcPBM and 63 ChIP-seq data for human transcription factors. Scrutinizing the results of the individual data sets, we find several cases where a PWM model neglecting dependencies between binding site positions already yields a decent prediction performance. However, for a considerable fraction of data sets, the improvement gained by models capturing dependencies between adjacent and non-adjacent positions is substantial.

Subsequently, we focus on ChIP-seq data sets for those transcription factors with the greatest improvements in prediction performance using Slim models and further investigate their dependency structures using dependency logos. In Figure 1, we show three examples of dependency logos based on predictions of Slim models. For Nfe2, we observe heterogeneities caused by two different, mixed motifs, where the first is an E-box-like (CACGTG) motif and the second is the expected Nfe2 motif with consensus TGCTGAGTCAY. For c-Jun, we find a flexible spacer between the two half sites with consensus TGA and TCA that has also been reported by Badis *et al.* [B$^+$09a] for Jundm2 in mouse using PBM data and by Mathelier and Wasserman [MW13] using TFFMs on ChIP-seq data for human Jund. For Nrsf, we find that only the top-scoring binding sites cover the complete Nrsf motif, whereas the majority of sequences under the ChIP-seq peaks (68%) contain only the left half site (CTGTCC). While a dependency of nucleotide conservation on ChIP enrichment of the Nrsf motif has been reported before [B$^+$09b], the



(a) Nfe2 (b) c-Jun (c) Nrsf

Figure 1: Dependency logos of binding sites predicted by the Slim model for ChIP-seq data sets.

clear distinction between two modes of Nrsf binding discovered using the Slim model is novel and might be related to the diverse complexes of Nrsf with other factors [Y+11].

In summary, we find that binding landscapes of transcription factors are highly complex and diverse, including secondary or multiple motifs, partial motifs, flexible binding modes, or dependencies between neighboring and non-neighboring positions. Some of these cases could also be handled by specialized models based on *a-priori* expert knowledge, e.g., spaced PWM models for c-Jun or hidden Markov model-like approaches for Nrsf. The strength of the proposed Slim models is their flexibility to adjust to all these dependency structures without requiring *a-priori* knowledge of dependency structures, while dependency logos allow for dissecting dependency structures *a-posteriori* by visual inspection.

Talk outline

In the first part of the talk, we will explain Slim models on a conceptual level. While Slim models have been designed for modeling DNA motifs, the general concept of soft feature selection in a local mixture model might be applicable to other bioinformatics problems as well. In the second part, we will focus on the results obtained on ChIP-seq data. We will briefly explain dependency logos and use these for visualizing several, representative dependency structures detected in transcription factor binding sites. Finally, we will show that dependency logos may also help to visually detect dependencies in other sequence data.

References

[B+09a] Gwenael Badis et al. Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 324(5935):1720–1723, 2009.

[B+09b] Alexander W. Bruce et al. Functional diversity for REST (NRSF) is defined by in vivo binding affinity hierarchies at the DNA sequence level. *Genome Research*, 19(6):994–1005, 2009.

[Bai11] Timothy L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[ENC12] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 09 2012.

[G+13] Jan Grau et al. A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21):e197, 2013.

[H+11] Peter Huggins et al. DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics*, 27(17):2361–2367, 2011.

[K+13] Ivan Kulakovskiy et al. From Binding Motifs In ChIP-seq Data To Improved Models Of Transcription Factor Binding Sites. *Journal of Bioinformatics and Computational Biology*, 11(01):1340004, 2013.

[KG15] Jens Keilwagen and Jan Grau. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Research*, 2015.

[M+11] Quaid Morris et al. Jury remains out on simple models of transcription factor specificity. *Nat Biotech*, 29(6):483–484, 06 2011.

[M+13] Fantine Mordelet et al. Stability selection for regression-based models of transcription factor–DNA binding specificity. *Bioinformatics*, 29(13):i117–i125, 2013.

[MW13] Anthony Mathelier and Wyeth W. Wasserman. The Next Generation of Transcription Factor Binding Site Prediction. *PLoS Comput Biol*, 9(9):e1003214, 09 2013.

[W+13] Matthew T. Weirauch et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, 31(2):126–134, February 2013.

[Y+11] Hong-Bing Yu et al. Coassembly of REST and its cofactors at sites of gene repression in embryonic stem cells. *Genome Research*, 21(8):1284–1293, 2011.

[ZS11] Yue Zhao and Gary D. Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology*, 29(6):480–483, June 2011.

PeerJ PrePrints

Computing and Visualizing Precision-Recall Curves and Receiver Operating Characteristic Curves for Soft-labeled and Hard-labeled Data

Ivo Grosse[1,2], Jan Grau[1] and Jens Keilwagen[3]
[1] *Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany*
[2] *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany*
[3] *Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany*
grosse@informatik.uni-halle.de

PeerJ PrePrints

Introduction

The assessment of classifier performance is of fundamental importance in many bioinformatics applications. For instance, measures of classification performance are used to select appropriate models for solving classification problems. Performance evaluation is also inevitable for demonstrating the utility of novel approaches. In general, it assists researchers in identifying the most promising approach for the classification problem at hand. This implies that the choice of appropriate performance measures may influence the results of downstream analyses.

For binary classification tasks, the receiver operating characteristic (ROC) curve and the area under this curve (AUC-ROC) are widely accepted as a general measure of classifier performance. In many bioinformatics applications, however, positive examples are substantially less abundant than negative examples, resulting in a highly imbalanced class ratio. For instance, the number of true donor splice sites is substantially smaller than the number of genomic sequences with central GT consensus, and the number of target genes of a microRNA is substantially smaller than the number of non-target genes. In such cases, the precision-recall (PR) curve and the area under this curve (AUC-PR) is better suited for comparing the performance of individual classifiers than the ROC curve and AUC-ROC [DBR$^+$05].

Often, the decision for the true class labels of a given data point is ambiguous and partly subjective. For instance, class labels may be based on an arbitrary threshold for some continuous measurement, e.g., fold changes of differentially expressed genes. Uncertain class labels may also arise from multiple, possibly contradictory, expert labelings. However, the decision for a specific labeling decisively influences classifier training and assessment. One solution to this problem is the transition from hard-labeling to soft-labeling, where each data point is assigned to both classes with a certain probability that reflects confidence in the labeling. For instance, Grau *et al.* [GPGK13] develop a schema for deriving soft-labels from peak statistics for ChIP-seq data, or Mihaljevic *et al.* [M$^+$14] determine soft-labels from expert labelings of interneurons. While soft-labeling has been used extensively for classifier training in the past, it has been neglected for classifier assessment [KGG14].

Computing empirical AUC-PR and AUC-ROC values from test data points requires interpolation between discrete supporting points corresponding to a series of classification thresholds. AUC-ROC can be computed by linear interpolation between the supporting points of the curve for hard-labeled and soft-labeled data. In contrast, Davis & Goadrich [DG06] show that for AUC-PR an interpolation along the true positives is more accurate than linear interpolation for hard-labeled data, while Boyd *et al.* [BEP13] and Keilwagen *et al.* [KGG14] propose a more fine-grained, continuous interpolation between the supporting points of the PR curve. Only the latter can also be used for soft-labeled data and weighted data in general.

We make this interpolation available to the scientific community in the R package PRROC [GGK15], which is available from CRAN and may be used to compute and visualize PR and ROC curves.

Results

To illustrate the efficacy of the developed method, we investigate the influence of soft-labeled test data on classifier performance. To this end, we compare the classifier performance of published classifiers using AUC-PR on hard-labeled and soft-labeled test data for predicting transcription factor binding affinities.

We perform a reassessment of classifiers from Weirauch *et al.* [WCN+13], who evaluate the performance of classifiers for 66 protein binding microarray (PBM) data sets. PBMs measure the *in-vitro* binding affinity of transcription factors to DNA sequences using microarrays in an unbiased manner, where double-stranded probe sequences are chosen such that they contain all k-mers up to a given k with identical frequency. The goal of that study was to assess different classifiers for their ability to distinguish bound from unbound probes and for the correspondence of their classification scores to measured microarray intensity values.

Weirauch *et al.* introduce a hard labeling based on the intensity values for all probes sequences in each of the 66 experiments. For each individual experiment, they define the threshold separating foreground and background data points. Based on this labeling, they compare classifiers using different performance measures including the mean AUC-ROC over all experiments.

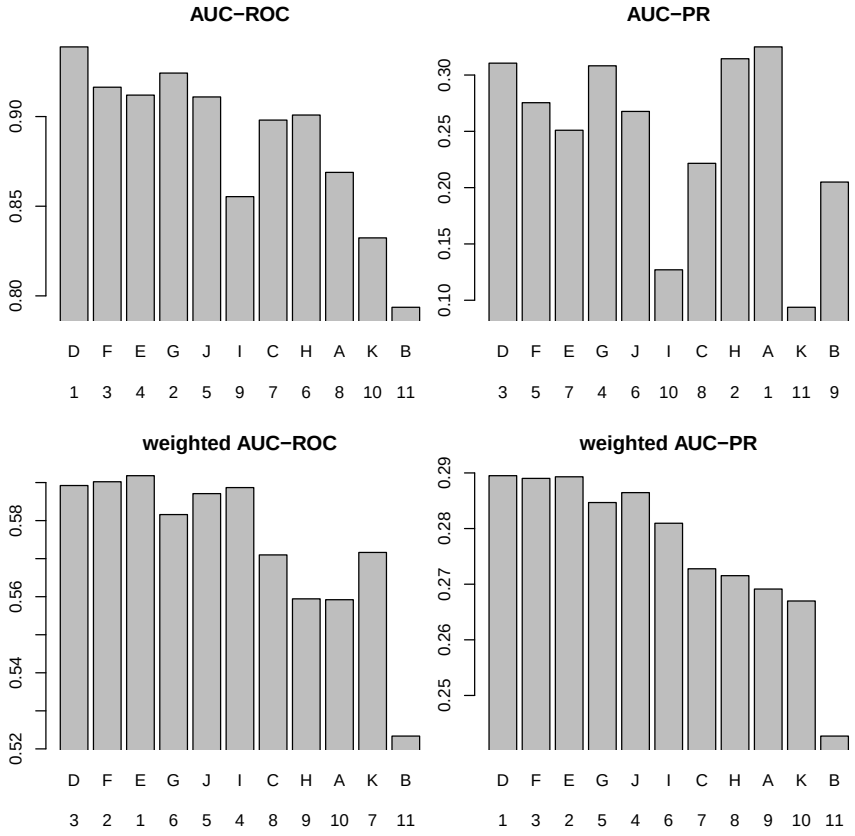


Figure 1: Mean results for AUC-ROC and AUC-PR on PBM data sets using hard-labeled or soft-labeled (i.e., weighted) test data. Letters (A,B,...,K) on the abscissa indicate the team names of approaches in the original publication of Weirauch *et al.* [WCN+13] and appear in the order of the original ranking. Rankings according to the different performance measures are shown below the team names, while the mean values for AUC-ROC and AUC-PR are depicted on the ordinate.

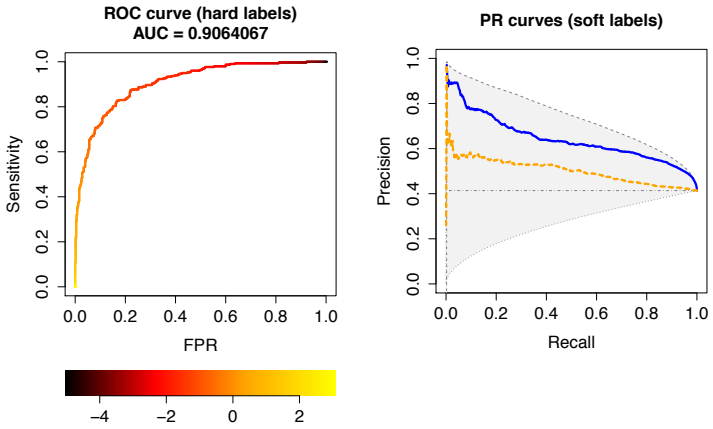


Figure 2: Plots of ROC (left) and PR (right) curves generated by PRROC. For the ROC curve, we consider hard-labeled data and show the plotting variant with a color scale that indicates classification thresholds yielding the points on the curve. For the PR curve, we consider soft-labeled data and show a comparative plot for two classifiers as solid blue and dashed orange lines. We also include the maximal and minimal possible curves and the curve of a random classifier for the given soft-labels.

In Figure 1, we compare the mean AUC-ROC, the mean AUC-PR, and the corresponding rankings for hard-labeled and soft-labeled test data. In the hard-labeled case, we take the class labels suggested by Weirauch *et al.* [WCN+13]. We find that the rankings for both mean AUC-ROC and mean AUC-PR change considerably when considering soft-labeled test data instead of less informative hard-labeled test data. Focusing on the mean AUC-PR, we find that the ranking obtained by AUC-PR using soft-labeled test data are in better accordance to the original ranking of Weirauch *et al.* than the ranking using hard-labeled test data.

PRROC R-package

We have developed a user-friendly and well-documented R package called PRROC [GGK15], which allows for computing PR and ROC curves as well as the areas under these curves for soft-labeled and hard-labeled data. Optionally, PRROC also computes curves and AUC values for the optimal, the worst, and the random classifier as a reference. These references are particularly useful for (i) PR curves and (ii) ROC and PR curves in case of soft-labeled data, where the minimum and maximum AUC may differ from 0 and 1, respectively. In addition, PRROC allows for visualizing PR and ROC curves as exemplarily shown in Figure 2. PRROC is available from CRAN (`http://cran.r-project.org/web/packages/PRROC/index.html`) and provides R documentation files and a vignette.

Talk outline

In the talk, we will first motivate why appropriate performance measures are important for classification problems in bioinformatics and why these should be chosen in a problem-specific manner. Second, we introduce AUC-PR as a useful performance measure for problems with highly imbalanced class ratios, which are prevalent in bioinformatics. Third, we will provide examples for bioinformatics applications that may profit from performance evaluation using soft-labels. Finally, we will show how researchers can use the PRROC R-package to evaluate classifier performance for soft-labeled and hard-labeled test data, and to produce publication-quality plots of PR and ROC curves using PRROC.

PeerJ PrePrints

References

[BEP13] Kendrick Boyd, Kevin H. Eng, and C. David Page. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *LNCS*, pages 451–466. Springer Berlin Heidelberg, 2013.

[DBR+05] Jesse Davis, Elizabeth Burnside, Raghu Ramakrishnan, Vitor Santos Costa, and Jude Shavlik. View learning for statistical relational learning: With an application to mammography. In *Proceeding of the 19th International Joint Conference on Artificial Intelligence*, pages 677–683, 2005.

[DG06] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, New York, NY, USA, 2006. ACM.

[GGK15] Jan Grau, Ivo Grosse, and Jens Keilwagen. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, 2015.

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

[KGG14] Jens Keilwagen, Ivo Grosse, and Jan Grau. Area under Precision-Recall Curves for Weighted and Unweighted Data. *PLoS ONE*, 9(3):e92209, 03 2014.

[M+14] Bojan Mihaljevic et al. Multi-dimensional classification of GABAergic interneurons with Bayesian network-modeled label uncertainty. *Frontiers in Computational Neuroscience*, 8(150), 2014.

[WCN+13] Matthew T. Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, DREAM consortium, Harmen J. Bussemaker, Quaid D. Morris, Martha L. Bulyk, Gustavo Stolovitzky, and Timothy R. Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31:126–134, 2013.

PeerJ PrePrints

Natural genetic variation impacts expression levels of coding, non-coding and antisense transcripts in fission yeast

Mathieu Clément-Ziza[1,2], Francesc X. Marsellach[3], Sandra Codlin[3], Manos A. Papadakis[4], Susanne Reinhardt[1], Maria Rodriguez-Lopez[3], Stuart Martin[3], Samuel Marguerat[3], Alexander Schmidt[5], Eunhye Lee[3], Christopher T. Workman[4], Jürg Bähler[3] , and Andreas Beyer [1,2]
[1] Biotechnology Centre, Technische Universität Dresden, Germany. [2] CECAD, University of Cologne, Köln, Germany. [3] University College London, London, United Kingdom. [4] Center for Biological Sequence Analysis, Technical university of Denmark, Denmark. [5] Biozentrum, University of Basel, Switzerland
mathieu.clement-ziza@uni-koeln.de

PeerJ PrePrints

Abstract

Our current understanding of how natural genetic variation affects gene expression beyond well-annotated coding genes is still limited. The use of deep sequencing technologies for the study of expression quantitative trait loci (eQTLs) has the potential to close this gap. Here, we generated the first recombinant strain library for fission yeast and conducted an RNA-seq-based QTL study of the coding, non-coding, and antisense transcriptomes. We show that the frequency of distal effects (*trans*-eQTLs) greatly exceeds the number of local effects (*cis*-eQTLs) and that non-coding RNAs are as likely to be affected by eQTLs as protein-coding RNAs. We identified a genetic variation of *swc5* that modifies the levels of 871 RNAs, with effects on both sense and antisense transcription, and show that this effect most likely goes through a compromised deposition of the histone variant H2A.Z. The strains, methods, and datasets generated here provide a rich resource for future studies [CZMC$^+$14].

Introduction

Variation in gene expression, which in turn is often caused by natural genetic variation, is a major factor causing intra-species phenotypic differences. Hence, investigating the influence of genetic variation on gene expression has been the focus of intense research. The identification of expression quantitative trait loci (eQTLs), that is, genomic regions that are linked with the expression of a specific transcript, has primarily been conducted using DNA microarrays [BYCK02]. High-throughput sequencing of cDNA (RNA-seq) has great potential to provide qualitatively and quantitatively new insights beyond mere mRNA quantification.

Previous RNA-seq based eQTL studies have focused on measuring new traits, and have been limited to the detection of local eQTLs, so called *cis*-eQTLs [LHW$^+$11, MP11, PMP$^+$10], or have identified a only a relatively small proportion of distant eQTLs [BMZ$^+$13]. *cis*-eQTLs are located at or close to the genes whose expression they directly affect; while *trans*-eQTLs are remote from the genes whose expression they affect. Further, sequence variation information contained in the RNA-seq data has not been exploited in the framework of eQTL mapping.

Here, we have conducted an expression QTL study characterized by a design enabling a high statistical power for association detection and by a broad investigation of pervasive expression beyond well-annotated coding genes (i.e. non-coding and antisense transcripts). The high statistical power contributed to the improved discovery of *trans*-eQTLs, suggesting that previous studies may have been overestimating the fraction of *cis*-eQTLs. First, we generated a recombinant strain library for fission yeast (*Schizosaccharomyces pombe*) suitable for powerful QTL studies, which was subsequently subjected to high-resolution measurements of growth kinetics and strand-specific RNA-seq. Whereas microarray probes rely on a fixed reference genome, RNA-seq allows for the individualized quantifica-

tion of transcripts taking genomic variation into account. We show that our approach, which explicitly includes individual genomes, reduces the potential for false-positive eQTLs. Further, because RNA-seq measures the actual transcript sequences of a given strain, it can also be used for genotyping the strain library. We developed a computational framework for the robust genotyping of recombinant strains using RNA-seq data, which eliminates the need for separate genotyping experiments. Finally, RNA-seq makes no assumptions about the structure of genomic features. In the context of QTL studies, it can thus be used to identify genetic variants affecting non-annotated features. Here, we present a striking example of a variation affecting antisense transcription of hundreds of *S. pombe* genes detected in this study.

Results and methods

Generation and phenotyping of a recombinant fission yeast strain library

We generated the first recombinant strain library for *Schizosaccharomyces pombe* suitable for QTL studies. In order to enable the detection of association at a high statistical power, we selected closely related parental strains to reduce the genetic complexity of the library (0.05% divergence, comparable to the average divergence between two humans). This cross was subsequently subjected to high-resolution measurements of growth kinetics and deep strand-specific RNA-seq (average effective depth 37.5x).

Genotyping of the strain library by RNA-seq

eQTL studies require both the genotypes and the expression profiles of each individual of the studied population. We developed a strategy enabling the genotyping of a recombinant strain library through RNA-seq. Thus, separate genotyping experiments of the segregant strains were not needed. First we sequenced the genome of the parental strains a great depth in order to detect potentially all genomic variants. Then we used RNAseq data of the segregants to detect genomic variation at the sites polymorphic when comparing the progenitors. On average half of the sites could be directly genotyped after conservative filtering. It was sufficient to identify haplotype blocks and thus infer genotypes at the remaining sites. This lead to the genotyping of the whole library at 4,481 sites.

Accounting for individual genomes improves transcript quantification

In microarray-based expression studies, sequence variation in probe regions can affect the hybridization efficiency. Because this leads to an allele-specific signal bias, sequence variation can inflate the number of false *cis*-eQTL calls [ATL$^+$07]. Notably, RNA-seq studies are neither immune to such artifacts [DMP$^+$09] as transcript quantification usually involves the mapping of sequence reads to a reference genome. In this study, gene expression quantification was performed by aligning reads to strain-specific genomes in order to minimize this bias. Using both simulations and real data, we compared this strategy to reference genome mapping. Results show that aligning RNA-seq data against individualized genomes marginally improves transcript quantification, while ignoring individual sequence variation can inflate the number of falsely detected *cis*-eQTLs.

PeerJ PrePrints

trans-eQTLs greatly exceed cis-eQTLs in abundance

After mapping the QTLs using a Random Forest based approach [MASB10, PCZL+13], one of the most surprising results of this study was the small fraction of *cis*-eQTL that were detected. It is generally assumed that *cis*-eQTLs can be detected more easily than *trans*-eQTLs [ASWB13, HLB+11, SMD+03]: (i) direct effects are stronger that distant indirect ones, and (ii) searching for *trans* linkages involves testing a much larger number of hypotheses. Strikingly we detected a much higher fraction of trans-eQTLs (˜90%) than *cis*-eQTL. We attribute this to the high statistical power of our study, which could be due to several factors: the genetic similarity of the parental strains reducing the complexity, the use of deep sequencing reducing trait noise, and/or the advanced methods we used to analyze these data.

Non-coding and expression are strongly affected by genetic variation

RNA-seq enables the quantification of entire transcriptomes, including non-coding RNAs (ncRNAs). The high sequencing depth used here and the extensive annotation of the fission yeast genome enabled us to quantify transcript levels for 1,428 annotated ncRNAs. Thus, this analysis presents the first comparative eQTL mapping for coding versus non-coding transcript levels at a genomic scale. We showed that that the expression of non-coding RNAs is at least as much affected by genetic variation as the expression of protein-coding RNAs. To further investigate the importance of non-coding RNAs as effectors of eQTLs, we predicted the most likely causal gene for each eQTL. Our result suggests that non-coding RNAs substantially contribute as effectors of the genetic variation of gene expression.

A frameshift in swc5 causes major eQTL hotspot, reduces H2A.Z deposition increase antisense transcription

We identified an eQTL hotspot (locus regulating numerous genes) affecting the sense expression of 817 genes and the anti-sense expression of 1,384 traits. This QTL hotspot shows more widespread gene expression effects than any other hotspot reported so far. Because of its extraordinary strength, we wanted to unravel its molecular basis. We identified a frame-shift polymorphism in the gene *swc5* as being the molecular regulator at this locus. Swc5 is a component of the Swr1 protein complex controlling the chromosomal deposition of the histone variant H2A.Z. H2A.Z has been associated with the control of antisense transcription if fission yeast [ZFZ+09]. We showed that the effect of *swc5* hotspot most likely goes through a compromised deposition of the histone variant H2A.Z, which consequently leads to an increase of read-through antisense transcription. We performed numerous experiments and analyses that all corroborated this hypothesis. Notably we studied expression changes in strains deleted for *swc5*, and we analyzed the H2A.Z occupancy via ChIP-seq.

Conclusion

Several methodological aspects have been developed in this study, for instance RNA-seq based genotyping or strain specific genome mapping. Moreover, the high statistical power to detect eQTLs characterizing this study led to interesting findings regarding the genetic control of non-coding expression and the relative importance of *trans* effect. The detailed experimental and analytic validation on one of the casual genes (*swc5*) offers new insights on how a QTL could modulate its target genes. This study has been published in *Molecular System Biology* [CZMC+14].

PeerJ PrePrints

Presentation outline

The presentation will first motivate the need of studying the genetic basis of molecular traits. Then the concepts of eQTL mapping will be presented. The main part of the presentation will focus on both methodological aspects (RNA-seq based genotyping or random forest based QTL mapping), and on the result highlights (the importance of the regulation of non-coding RNA, the proportion of *cis*/*trans*). Finally the *swc5* eQTL hotspot will be briefly presented.

References

[ASWB13] M. Ackermann, W. Sikora-Wohlfeld, and A. Beyer. Impact of Natural Genetic Variation on Gene Expression Dynamics. *PLoS Genet*, 9(6):e1003514, June 2013.

[ATL+07] Rudi Alberts, Peter Terpstra, Yang Li, Rainer Breitling, Jan-Peter Nap, and Ritsert C. Jansen. Sequence Polymorphisms Cause Many False cis eQTLs. *PLoS ONE*, 2(7):e622, July 2007.

[BMZ+13] A. Battle, S. Mostafavi, X. Zhu, J. B Potash, C. Weissman, M. M .and McCormick, C. D Haudenschild, K. B Beckman, J. Shi, R. Mei, A. E Urban, S. B Montgomery, Douglas F Levinson, and D. Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, October 2013.

[BYCK02] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.)*, 296(5568):752–755, April 2002.

[CZMC+14] M. Clément-Ziza, F. X. Marsellach, S. Codlin, M. A. Papadakis, S. Reinhardt, M. Rodriguez-Lopez, S. Martin, S. Marguerat, A. Schmidt, E. Lee, C. T. Workman, J. Bahler, and A. Beyer. Natural genetic variation impacts expression levels of coding, noncoding, and antisense transcripts in fission yeast. *Molecular Systems Biology*, 10(11):764, November 2014.

[DMP+09] J. F Degner, J. C Marioni, A. A Pai, J. K Pickrell, E. Nkadori, Y. Gilad, and J. K Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics (Oxford, England)*, 25(24):3207–3212, December 2009.

[HLB+11] B. Holloway, S. Luck, M. Beatty, J-A Rafalski, and B. Li. Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *BMC Genomics*, 12(1):336, June 2011.

[LHW+11] E. Lalonde, K. C.H. Ha, Z. Wang, A. Bemmo, C. L. Kleinman, T. Kwan, T. Pastinen, and J. Majewski. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Research*, 21(4):545–554, April 2011.

[MASB10] J.J. Michaelson, R. Alberts, K. Schughart, and A. Beyer. Data-driven assessment of eQTL mapping methods. *BMC Genomics*, 11(1):502, 2010.

[MP11] Jacek Majewski and Tomi Pastinen. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics*, 27(2):72–79, February 2011.

[PCZL+13] P. Picotti, M. Clément-Ziza, H. Lam, D. S. Campbell, A. Schmidt, E. W. Deutsch, H. Rst, Z. Sun, O. Rinner, L. Reiter, Q. Shen, J. J. Michaelson, A. Frei, S. Alberti, U. Kusebauch, B. Wollscheid, R. L. Moritz, A. Beyer, and R. Aebersold. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature*, 494(7436):266–270, February 2013.

[PMP+10] JK. Pickrell, JC. Marioni, AA. Pai, JF. Degner, BE. Engelhardt, E. Nkadori, J-B. Veyrieras, M. Stephens, Y. Gilad, and JK. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, March 2010.

[SMD+03] E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, March 2003.

[ZFZ+09] M Zofall, T. Fischer, K. Zhang, M. Zhou, B. Cui, T. D. Veenstra, and S.I.S. Grewal. Histone H2A.Z cooperates with RNAi and heterochromatin factors to suppress antisense RNAs. *Nature*, 461(7262):419–422, August 2009.

PeerJ PrePrints

Integrating Sequence and Structure Information for Efficient Retrieval and Alignment of Flexible Protein Binding Sites

Stefan Bietz and Matthias Rarey
Center for Bioinformatics, University of Hamburg, Hamburg, 20146, Germany
rarey@zbh.uni-hamburg.de

The consideration of protein flexibility is long known to be one of the most challenging aspects in computational structural biology. Experimental structures are often used to construct flexible protein models or serve as a reference for flexibility predicting techniques. Therefore, the steadily increasing amount of structural data further improves the fundamental basis of these techniques. However, due to inconsistent annotation or structural deviations in the experimental data, alignment techniques are generally required as an essential prepressing step for the usage of multiple experimental protein structures. In principle, various sequence- and structure-based approach have already been developed which can be applied in these [FRCC07, KSK11]. However, their applicability depends on the particular application scenario. Many structure-based approaches like molecular docking, pharmacophore generation, or the investigation of enzymatic mechanisms focus on protein binding sites and neglect the rest of the protein structure for efficiency reasons. For these applications, it is most relevant that the alignment of the active site is highly reliable. Furthermore, purely sequence-based alignment techniques are not always applicable if the binding site is located at a subunit interface of an oligomeric protein, as in these cases, the assignment of corresponding subunits is not necessarily unambiguous. Structure-based alignment methods mostly target the identification of geometrically conserved motifs which complicates the identification of analogous protein regions exhibiting structural flexibility.

We recently introduced ASCONA [BR15a], an automated approach for the detection and alignment of protein binding site conformations. While most other alignment techniques deal with the generation of structure or sequence alignments of rather distantly related proteins, ASCONA puts the accurate detection of highly deviating binding site conformations into focus. Given an arbitrarily defined binding site of a query structure, ASCONA locates all occurrences of the respective query in a target structure and generates a residue-wise mapping in form of a sequence alignment. It also facilitates the generation of multiple alignments of a certain query in case of an oligomeric target structure.

The underlying algorithm is based on a partition of the query binding site into a set of short peptide fragments which are searched in the target sequence using an efficient approximate string matching algorithm. The fragment hits are recombined on the basis of a two-step fragment assembly approach and a geometry measure that analyses the distance and relative orientation of the fragment hits. Since the typical application scenario assumes a high sequence similarity of query and target structures, the fragment matching step can be set up quite strictly, which results in low rate of random (false positive) fragment matches. In turn, this allows for applying a tolerant geometry measure during the fragment assembly and thus facilitates an accurate detection of binding sites with highly deviating conformations. ASCONA was evaluated on the Astex Non-native dataset [VMH$^+$08] and proved to correctly align all contained binding sites including those with considerable structural deviations. A major advantage of ASCONA is that it only needs to search for the protein region of interest, e.g. a ligand binding site or a protein-protein interface, and thus achieves considerably low computation times. For instance, the alignment of a structure from the Astex Non-native dataset took on average 4 milliseconds.

Besides details on its algorithmic background and the evaluation experiments demonstrating its general functionality, we will present further information on sensible application scenarios. For instance, we developed a server for collecting protein binding site ensembles from the PDB [BR15b]. Starting with a user defined query, the search initially extracts structure candidates from a database that has been specially geared to this purpose. In a second step, ASCONA is used to detect appropriate binding sites within the set of candidates. This step highly benefits from ASCONA's accuracy and efficient runtime behavior. The remaining structures can be further filtered to adapt the set of identified binding

PeerJ PrePrints

site conformations to the user's requirements. This can, e.g., incorporate the application of RMSD thresholds, mutation rate constraints, or the selection of diverse conformations. These filters also depend on an accurate alignment of the query binding site and the ensemble candidates. Finally, the selected conformations are being superimposed on the basis of a common rigid region.

In summary, ASCONA is a perfectly suited tool for the collection and automatic preprocessing of alternative protein binding site conformations and can support any application that relies on an accurate mapping of the residues in the protein binding site.

References

[BR15a] Stefan Bietz and Matthias Rarey. ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations. *Journal of Chemical Information and Modeling*, 2015. (accepted).

[BR15b] Stefan Bietz and Matthias Rarey. Efficient Search of Experimentally Derived Structures for the Selective Compilation of Protein Binding Site Ensembles. 2015. (in preparation).

[FRCC07] Piero Fariselli, Ivan Rossi, Emidio Capriotti, and Rita Casadio. The WWWH of remote homolog detection: The state of the art. *Briefings in bioinformatics*, 8(2):78–87, 2007.

[KSK11] Singarevelu Kalaimathy, Ramanathan Sowdhamini, and Karuppiah Kanagarajadurai. Critical assessment of structure-based sequence alignment methods at distant relationships. *Briefings in bioinformatics*, 12(2):163–175, 2011.

[VMH+08] Marcel L Verdonk, Paul N Mortenson, Richard J Hall, Michael J Hartshorn, and Christopher W Murray. Protein- ligand Docking against Non-native Protein Conformers. *Journal of Chemical Information and Modeling*, 48(11):2214–2225, 2008.

PeerJ PrePrints

Fast alignment-free sequence comparison using spaced-word frequencies

Chris-André Leimeister, Marcus Boden,
Sebastian Lindner, Sebastian Horwege, Burkhard Morgenstern
University of Göttingen, Institute of Microbiology and Genetics,
Department of Bioinformatics, Goldschmidtstr. 1, 37073 Göttingen, Germany
bmorgen@gwdg.de

Sequence alignment is traditionally the first step in DNA and protein sequence analysis. With the amount of sequence data that are now available, however, pairwise or multiple alignment has become too slow in many applications. Therefore, alignment-free methods are increasingly used for genome comparison and phylogeny reconstruction, and the development of such methods has become a very active area of bioinformatics [Vin14]. While alignment-free methods are generally less accurate than alignment-based approaches, they are much faster since they run in linear time. Most alignment-free algorithms work by comparing the *word composition* of sequences. Sequences are represented by *word-frequency vectors*, and standard distance measures on vector spaces can be applied to calculate a pairwise distance matrix for a set of input sequences [HRR06, CHL+09, VCF+12, SJWK09]. Phylogenetic trees can then be calculated from these distance matrices with the usual distance-based methods for phylogeny reconstruction. A certain drawback of these word-based methods is the fact that word occurrences at adjacent sequence positions are far from independent.

Database search programs such as *BLAST* [AGM+90] originally used *word matches* of a fixed length k as *seeds* to search for local homologies. Here, the seed length k is a trade-off between *sensitivity* and *speed*. It has been shown that the sensitivity and speed of these programs can be substantially improved if *spaced seeds* – *i.e.* word matches with possible mismatches at certain pre-defined *mismatch positions* – are used instead of *contiguous* word matches as used in the original version of *BLAST* [MTL02]. Considerable efforts have been made since then, to find suitable patterns for this *spaced-seed* approach, see *e.g.* [BBV04, KNR06, Bro08, IIB11].

Inspired by these approaches, we previously proposed to use *spaced words* for alignment-free sequence comparison, *i.e.* words containing *wildcard* characters at fixed positions, according to an underlying *pattern* P of *match* and *don't care* positions [BSH+13]. The first version of our approach used one single pattern P: for a given set of input DNA or protein sequences and a pattern P, we calculated pairwise distances based on the spaced-word frequency vectors of the sequences with respect to P. A certain draw-back of this original *single-pattern* approach was the necessity to select one specific pattern P of *match* and *don't care* positions, since the results of this method strongly depend on the selected pattern.

In a subsequent paper [LBH+14], we used a *hashing* algorithm to compare the spaced-word composition of sequences that was much more efficient than the tree-based algorithm that we used in the previous implementation. This way, we were able to extend our approach to using sets $\mathcal{P} = \{P_1, \ldots, P_m\}$ of randomly generated patterns P_i of a fixed length and number of *match* positions, instead of a single pattern P. (Multiple patterns of *match* and *don't care* positions have also been proposed to generate *spaced seeds* for database searching [LMKT03].) In this *multiple-pattern* version of our approach, spaced-word frequencies are then calculated and compared with respect to *all* patterns in the set $\mathcal{P}$; we define the distance between two sequences as the *average* distance over all distance values obtained with the individual patterns $P_i \in \mathcal{P}$ that are calculated as in our previous *single-pattern* approach.

As in our previous paper, we evaluated this *multiple-pattern approach* by applying it to phylogeny analysis. We tested two different approaches to calculate pairwise distances between the input sequences based on their (multiple) spaced-word-frequencies, namely the *Euclidean* distance and the *Jensen-Shannon* distance [Lin91]. The resulting distance matrices were used as input for *Neighbour Joining* [SN87] to generate trees, and we compared the resulting tree topologies to trusted reference topologies using the *Robinson-Foulds* distance [RF81]. As benchmark data sets, we used simulated and real-world

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1352v1 | CC-BY 4.0 Open Access | rec: 7 Sep 2015, publ: 7 Sep 2015

DNA and protein sequences.

In our first paper, we had shown that the *single-pattern* version of our *spaced-words* leads to slightly better trees than the same approach used with *contiguous words* [BSH+13]. In [LBH+14], we could show that our new *multiple-pattern* approach produces much better phylogenies than the previously implemented *single-pattern* approach and is also superior to established alignment-free methods that are based on *contiguous* words. On some data sets, the quality of our results was even comparable to trees that were obtained with traditional alignment-based approaches.

Also, we showed empirically that distance values calculated with our *multiple-pattern* program are statistically more stable than distances based on the previous *single-pattern* approach which were, again, more stable than distances based on the frequencies of *contiguous* words. In a more recent paper [MZHL15], we studied the statistical behaviour of our spaced-word-based distance functions in detail and showed analytically why spaced-word-based distances are statistically more stable than distances calculated from contiguous words and why, in turn, the new *multiple-pattern* version of *spaced words* is more stable than the previous *single-pattern* approach.

Our software is freely available as source code. In addition, we provide a user-friendly WWW interface that is described in [HLB+14]. Source code and WWW interface are available through *Göttingen Bioinformatics Compute Server (GOBICS)* at

`http://spaced.gobics.de/`

References

[AGM+90] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene M. Myers, and David J. Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215:403–410, 1990.

[BBV04] Brona Brejova, Daniel G. Brown, and Tomas Vinar. Optimal Spaced Seeds for Homologous Coding Regions. *Journal of Bioinformatics and Computational Biology*, 1:595–610, 2004. Early version appeared in CPM 2003.

[Bro08] Daniel G. Brown. *Bioinformatics Algorithms: Techniques and Applications*, chapter A survey of seeding for sequence alignment, pages 126–152. Wiley-Interscience, New York, Fev. 2008.

[BSH+13] Marcus Boden, Martin Schöneich, Sebastian Horwege, Sebastian Lindner, Chris-André Leimeister, and Burkhard Morgenstern. Alignment-free sequence comparison with spaced *k*-mers. In Tim Beißbarth, Martin Kollmar, Andreas Leha, Burkhard Morgenstern, Anne-Kathrin Schultz, Stephan Waack, and Edgar Wingender, editors, *German Conference on Bioinformatics 2013*, volume 34 of *OpenAccess Series in Informatics (OASIcs)*, pages 24–34, Dagstuhl, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[CHL+09] Benny Chor, David Horn, Yaron Levy, Nick Goldman, and Tim Massingham. Genomic DNA k-mer spectra: models and modalities. *Genome Biology*, 10:R108, 2009.

[HLB+14] Sebastian Horwege, Sebastian Lindner, Marcus Boden, Klaus Hatje, Martin Kollmar, Chris-André Leimeister, and Burkhard Morgenstern. *Spaced words* and *kmacs*: fast alignment-free sequence comparison based on inexact word matches. *Nucleic Acids Research*, 42:W7–W11, 2014.

[HRR06] Michael Höhl, Isidore Rigoutsos, and Mark A. Ragan. Pattern-Based Phylogenetic Distance Estimation and Tree Reconstruction. *Evolutionary Bioinformatics Online*, 2:359–375, 2006.

[IIB11] Lucian Ilie, Silvana Ilie, and Anahita M. Bigvand. SpEED: fast computation of sensitive spaced seeds. *Bioinformatics*, 27:2433–2434, 2011.

[KNR06] Gregory Kucherov, Laurent Noé, and Mikhail Roytberg. A unifying framework for seed sensitivity and its application to subset seeds. *Journal of Bioinformatics and Computational Biology*, 4:553–569, 2006.

[LBH+14] Chris-André Leimeister, Marcus Boden, Sebastian Horwege, Sebastian Lindner, and Burkhard Morgenstern. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30:1991–1999, 2014.

PeerJ PrePrints

[Lin91] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37:145–151, 1991.

[LMKT03] Ming Li, Bin Ma, Derek Kisman, and John Tromp. PatternHunter II: Highly Sensitive and Fast Homology Search. *Genome Informatics*, 14:164–175, 2003.

[MTL02] Bin Ma, John Tromp, and Ming Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18:440–445, 2002.

[MZHL15] Burkhard Morgenstern, Bingyao Zhu, Sebastian Horwege, and Chris-André Leimeister. Estimating Evolutionary Distances between Genomic Sequences from Spaced-Word Matches. *Algorithms for Molecular Biology*, 10:5, 2015.

[RF81] DF Robinson and LR Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[SJWK09] Gregory E. Sims, Se-Ran Jun, Guohong A. Wu, and Sung-Hou Kim. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106:2677–2682, 2009.

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

[VCF+12] Susana Vinga, Alexandra M. Carvalho, Alexandre P. Francisco, Luís M. S. Russo, and Jonas S. Almeida. Pattern matching through Chaos Game Representation: bridging numerical and discrete data structures for biological sequence analysis. *Algorithms for Molecular Biology*, 7:10, 2012.

[Vin14] Susana Vinga. Editorial: Alignment-free methods in computational biology. *Briefings in Bioinformatics*, 15:341–342, 2014.

PeerJ PrePrints

Ultra-fast functional classification of short reads using UProC with Pfam and KEGG

Manuel Landesfeind, Robin Martinjak, Heiner Klingenberg and Peter Meinicke
Department of Bioinformatics, Institute of Microbiology and Genetics, University of Göttingen
peter@gobics.de

As introduced in [Mei15] and [LM14a] we here present a novel tool UProC for large-scale sequence classification and show its application to functional analysis of metagenomic and transcriptomic data.

The functional and metabolic characterization of organisms and organism communities based on massive sequencing is central in genome and metagenome studies. With the current next-generation sequencing techniques the number of reads per sample has dramatically increased while in comparison to Sanger-sequencing the average read-length has substantially decreased. Because the vast amount of short read data renders a classical BLAST-based analysis infeasible, novel tools have to be developed to cope with the already existing computational bottleneck. In metagenomics and metatranscriptomics the functional classification of sequencing reads based on assignments to annotated protein sequences or families is usually the computationally most expensive task. Using a sensitive BLASTX search [AGM+90] against a large protein sequence database, the processing of millions of short reads on an actual desktop computer can take years and makes massive parallelization on large computer clusters inevitable. Using HMMER profile HMMs [Edd98] is about one order of magnitude faster but would be restricted to protein families that can be well represented by multiple sequence alignments. Therefore, we have developed the **U**ltra-fast **Pro**tein domain **C**lassification (UProC) tool which is about 10000 and 1000 times faster than BLASTX and HMMER3, respectively. The UProC algorithm features a novel sequence scoring approach that we refer to as "Mosaic Matching". Although, UProC has been designed to assign sequences to protein domain families we have also used it with the KEGG database of full-length gene families (KEGG Orthologs) and found it well-suitable to predict the functional repertoire of an organism from unassembled RNAseq reads [LM14a]. We have used UProC for the development of several tools that require the computation of functional profiles [KALM13, LM14b, AWDM15] and it makes up the core of our CoMet [LASM11] and CoMet-Universe [AKLM14] web servers. The UProC source code is available at `https://github.com/gobics/uproc` with the latest release package (including precompiled binaries for Windows) and databases provided at `http://uproc.gobics.de/`.

1 UProC algorithm and Pfam domain classification results

The UProC "Mosaic Matching" algorithm comprises several essential elements which involve the scoring and classification of protein sequences as well as the prior database construction. In this extended abstract we will only give a short overview of the most basic steps and would like to point the interested reader to the original publication [Mei15] for a full description.

The algorithm first extracts all oligopeptides ("words") of length 18 from a query protein sequence and for every word identifies the nearest neighbour in a sorted database array of labelled reference words. All residues of a nearest neighbour word are then scored with respect to their similarity to the query word using a position specific scoring matrix that has been optimized by a machine learning approach. Finally all position specific scores from reference words with the same protein family label are combined in a Mosaic Match to yield the final score of the sequence with respect to the particular family. If the score is above a length-dependent threshold a significant match is reported.

We evaluated UProC with the Pfam 24 database [FMT+10] on real and simulated metagenome data of varying complexity and read length, comparing it with HMMER3 and RPS-BLAST. We found that on the shortest read length (100 bp) UProC outperformed the profile-based tools in terms of sensitivity

PeerJ PrePrints

at a comparable specificity which was around 95% in all cases. The sensitivity of UProC varied from 88.9% on a human microbiome dataset to 68.5% on marine metagenome data, down to 50.1% on data from a microbial mat community. The corresponding sensitivity of HMMER (RPS-BLAST) for these datasets was 52.3 (48.5), 47.5 (44.8) and 42.8 (39.6) percent, respectively. The results indicate that the computationally more expensive profile methods which constitute the state-of-the-art for full length protein sequences might not be optimal for this kind of short read data. This has also been found in a recent study using transcriptomic data [ZSC13] which exhibited a substantial sensitivity loss of HMMER and other profile-based methods for the classification of protein domains in short reads. As expected, for increasingly longer reads, at some point HMMER becomes the most sensitive tool. For a good classification performance, UProC requires a large sequence database that covers much of the variation within different protein families. On the other hand, UProC does not require the protein families to be representable in terms of multiple alignments. At the UProC homepage we offer a precompiled database for a recent version of the KEGG orthologs [KG00] which are widely used for metabolic profiling in metagenomics and metatranscriptomics. An application of UProC to KEGG-based classification of short reads is reported in the following.

2 Using UProC with KEGG to predict functional repertoires from unassembled RNA-Seq data

In the annotation of *de novo* sequenced organisms the inference of potential gene functions is a fundamental step. If a genome sequence can be assembled at sufficient quality, for an automatic annotation of predicted genes, putative functions are usually identified using homology search techniques. Without the genomic sequence a *de novo* transcriptome assembly can be used to assess major parts of the functional repertoire where the achievable coverage strongly depends on the experimental setup and the organism under investigation. This strategy has been adopted as a valuable alternative for certain organisms which for example provide large or hybrid genomes. Although many tools have been proposed for *de novo* transcriptome assembly, the risk of misassemblies remains and also depends on the organism. In addition, the computational effort in terms of RAM storage requirements for the assembly can be demanding. Finally, the result of the analysis is highly dependent on several parameters, in particular on a suitable threshold for the homology search step, such as a BLAST E-value cut-off, which is necessary to decide on the presence or absence of a particular function.

In a recent study [LM14a] we have investigated to what degree it is possible to reconstruct the functional inventory of an organism using only unassembled transcriptome data. The short read data was directly mapped to KEGG functions by searching for homologies to the corresponding KEGG Ortholog families. For the evaluation we used a large RNA-Seq data set from *Arabidopsis thaliana* and removed all sequences of that organism and close relatives from the database. To obtain a reliable prediction on the presence of a function on the basis of short reads, it is important to evaluate the aggregated evidence that is generated by all reads showing similarity to reference sequences of the same family. The similarity scores calculated at the homology search step were combined in a family-specific evidence measure which was finally used for the prediction of the corresponding function. We found that over the whole range of possible functions the distribution of the evidence measure typically shows a bimodal distribution that reflects the dichotomy of strong and weak similarities with respect to different organisms in the database. This bimodality makes it possible to automatically adjust the prediction threshold using a mixture model for analysis of the evidence distribution.

Our results show a high sensitivity of up to 94 percent for the prediction of biomolecular functions in KEGG. The low false positive rate of 4 percent indicates that the automatic threshold calibration is highly effective even providing a better performance than prediction on the basis of a *de novo* transcriptome assembly. In our study we also compared the impact of different homology search tools, including several pairwise approaches and UProC. We found that the application of UProC provides the fastest solution and at the same time the highest detection performance (F1-measure) for this particular task.

PeerJ PrePrints

Thereby, the UProC memory requirements of approximately 16 GB RAM are clearly higher than with BLAST but much lower than for transcriptome assembly tools.

In metatranscriptomics, not only the functional characterization but also the phylogenetic classification of sequencing reads is required. Although, UProC can be used for taxonomic profiling of metagenomes by means of the Taxy-Pro mixture model [KALM13] for evaluation of protein domain counts, the taxonomic binning of reads is currently not possible. This is a clear advantage of BLASTX-based approaches (see e.g.[GS11, HMR+11]) that can provide both, functional and phylogenetic classification of single reads. Currently, we are working on a UProC version that integrates both kinds of classification.

References

[AGM+90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, Oct 1990.

[AKLM14] K. P. Aßhauer, H. Klingenberg, T. Lingner, and P. Meinicke. Exploring Neighborhoods in the Metagenome Universe. *International Journal of Molecular Sciences*, 15(7):12364–12378, July 2014.

[AWDM15] K. P. Asshauer, B. Wemheuer, R. Daniel, and P. Meinicke. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, May 2015.

[Edd98] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, January 1998.

[FMT+10] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res.*, 38:D211–222, Jan 2010.

[GS11] W. Gerlach and J. Stoye. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Research*, 39(14):e91–e91, August 2011.

[HMR+11] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–1560, September 2011.

[KALM13] H. Klingenberg, K. P. Aßhauer, T. Lingner, and P. Meinicke. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics*, 29(8):973–980, April 2013.

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

[LASM11] T. Lingner, K. P. Aßhauer, F. Schreiber, and P. Meinicke. CoMet–a web server for comparative functional profiling of metagenomes. *Nucleic Acids Research*, 39(Web Server issue):W518–523, July 2011.

[LM14a] M. Landesfeind and P. Meinicke. Predicting the functional repertoire of an organism from unassembled RNAseq data. *BMC Genomics*, 15(1):1003, November 2014.

[LM14b] Lingner, T. and Meinicke, P. Characterizing metagenomic novelty with unexplained protein domain hits. In *German Conference on Bioinformatics 2014*, GI-Edition : lecture notes in informatics, Proceedings, pages 69–78, 2014.

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

[ZSC13] Y. Zhang, Y. Sun, and J. R. Cole. A Sensitive and Accurate protein domain cLassification Tool (SALT) for short reads. *Bioinformatics*, 29(17):2103–2111, January 2013.

PeerJ PrePrints

Causal modeling of stroma-cancer cell communication

Julia C Engelmann[1], Claus Hellerbrand[2], Rainer Spang[1]
1 *Statistical Bioinformatics, University of Regensburg, Germany*
2 *Internal Medicine I, University Hospital Regensburg, Germany*
julia.engelmann@ur.de

Abstract: Molecular communication between stroma and cancer cells is well recognized to have a crucial role in carcinogenesis, tumor growth and cancer cell migration. From a systems biology perspective it links the molecular networks of both cell types. We have described a systems level analysis combining experimental and computational approaches for studying inter-cellular communication via secreted gene products. Our approach builds on statistical methodology for causal analysis based on a combination of reverse network engineering and causal effect estimation using genome scale observational data. In the context of molecular interactions between hepatic stellate cells (HSC) and hepatocellular carcinoma (HCC) cells, we predicted causal effects of HSC secreted gene products on tumor (HCC) gene expression. The many cause-effect pairs were then condensed to a small set of stromal factors which together cause the majority of gene expression changes observed in HCC cells with a Bayesian approach borrowed from Model-based Gene Set Analysis (MGSA). This resulted in a set of 10 secreted HSC gene products which together cause the majority of gene expression changes observed in HCC cells. The set of secreted stromal factors contained both known and unknown cancer promoting factors, including Placental Growth Factor (PGF) and Periostin (POSTN) as representatives of the former, and Pregnancy-Associated Plasma Protein A (PAPPA) as an example of the latter. We could show that PAPPA contributes to the activation of NFkB signaling. In clinical data, higher levels of PAPPA are linked to advanced stage HCC.

PeerJ PrePrints

1 Background

Cancer is a heterogeneous assembly of different cell types characterized, among others, by its composition and interactions of different cells. The basic building blocks of a cancer entity are epithelial cells, fibroblasts, vascular and inflammatory cells plus the extracellular matrix. Their interactions via (1) cell-cell contacts, (2) secreted factors like chemo- and cytokines, and (3) the modulation of the extracellular matrix are dynamic and influence cell proliferation, movement and differentiation [TC06].

Hepatocellular carcinoma (HCC) is one of the most prevalent and lethal malignant tumors worldwide. The major risk factor predisposing to HCC is hepatic cirrhosis. It arises through the activation of hepatic stellate cells (HSC), myofibroblast-like cells that are responsible for the excessive hepatic matrix deposition seen in chronically damaged livers. Moreover, HSCs infiltrate the stroma of liver tumors localizing around tumor sinusoids, fibrous septa, and capsules [WF14]. Conditioned medium collected from activated HSCs induces growth, migration and invasion of HCC cells in vitro. Furthermore, HSCs promote aggressive growth of HCC cells in experimental in vivo models [ZZY+11] and their presence predicts poor clinical outcome in HCC patients [JQF+09]. These data indicate that HSCs affect HCCs. Yet, the molecular mechanisms of this crosstalk are largely unknown.

Intra- and inter-cellular molecular mechanisms are typically studied using functional assays that involve the perturbation of the cellular systems. Unlike statistical associations in observational data, functional assays can directly distinguish between cause and effect. Their disadvantage is that they can be difficult to perform in high throughput. Recently, Maathuis and colleagues introduced a novel method to extract causal information from mere observational gene expression data [MKB09]. Their IDA ('Intervention calculus when the DAG is absent') algorithm combines local reverse network engineering using the PC-algorithm [SGS00] with causal effect estimation [Pea00, Pea03]. These virtual functional assays predict lists of genes that will change expression if the expression of a query gene was perturbed experimentally.

In [EAOR+15] we showed how functionally relevant secreted agents of stroma-tumor communication can be successfully predicted through a combination of novel experimental designs, causal network modeling, and data integration: Stromal hepatic stellate cells (HSC) from a set of human donors were

cultivated and the conditioned media were used to stimulate hepatocellular carcinoma cells (HCC). Gene expression was measured on the paired HSC and HCC cells before and after stimulation. With information on gene expression levels in both 'sender' and 'receiver' cells, we were able to infer which genes might play a role in communication of these two cell types. We used the IDA framework to predict the effects of virtual targeted interventions in HSCs on the expression of individual genes in stimulated HCCs. Finally, we integrated the large set of predicted pairs of causally interacting gene products to select the most important HSC secreted agents influencing cancer cell gene expression.

2 Results

Causal modeling approach

In our paper [EAOR+15], we used virtual targeted interventions by means of the IDA algorithm [MKB09] to identify gene products that mediate the communication of stroma and cancer cells. IDA consists of two steps. First, a partially directed network of regulatory interactions is constructed using the PC algorithm [SGS00]. Second, causal effects are estimated using Pearl's Do-calculus [Pea00]. To infer a potential causal effect of a stromal gene x on a cancer gene y, IDA needs the expression of y, x, and all genes x' that directly influence the expression of x in the regulatory network. Since stromal cells were in no contact to cancer cells in our experimental setting, the genes x' must be stromal genes as well. Hence it is sufficient to confine the reconstruction of a regulatory network to stromal genes only. For each of the cancer genes that changed expression upon conditioned media stimulation (False Discovery Rate < 0.001), we used IDA to screen for potential stromal genes that when perturbed in expression would have a strong effect on the respective cancer gene. Therefore we focused on secreted gene products as candidate stromal regulators. However, these genes are most likely regulated by non-secreted gene products which hence also need to be included into the network reconstruction. To limit the computational burden, we included the most highly and variably expressed genes across the stromal samples into the analysis, assuming that they would translate into abundant and variable amounts of protein. Since we were interested in cellular communication via secreted gene products, we confined the list of potential activators of cancer genes to only secreted stroma genes.

For each of the target cancer genes, secreted stroma genes were ranked by the effect size estimated by IDA. This procedure corresponds to ranking by the predicted causal effect in a virtual perturbation experiment: Gene-by-gene, all secreted stroma genes were virtually repressed by one standard unit and the expected change of the cancer gene was calculated. Performing the analysis on standardized data allows comparing effects across genes, and thus, the stromal gene with the strongest expected effect was ranked first, and so on. We applied IDA modeling in a sub-sampling approach, reporting causal effects only when they were insensitive to small perturbations of the data. The experimental and computational model set-up is depicted in Figure 1.

A small set of stroma-secreted proteins can activate cancer gene expression in concert.

Although all secreted HSC proteins have the potential to affect the expression of HCC genes, we postulate that a much smaller set of proteins is sufficient to activate HCCs. Thus in [EAOR+15] we aimed at identifying a small set of HSC genes that jointly account for the wide spectrum of expression changes in HCC cells observed in response to stimulation with HSC-CMs. We arranged the cause-effect pairs such that we obtained a list of potential HCC targets for each HSC cause. Since several HSC genes were predicted to affect multiple HCC genes, these lists overlapped. Model based Gene Set Analysis (MGSA) [BRG11] is an algorithm that aims at partially covering an input list of genes with as little Gene Ontology categories as possible. It balances the coverage with the number of categories needed.

PeerJ PrePrints

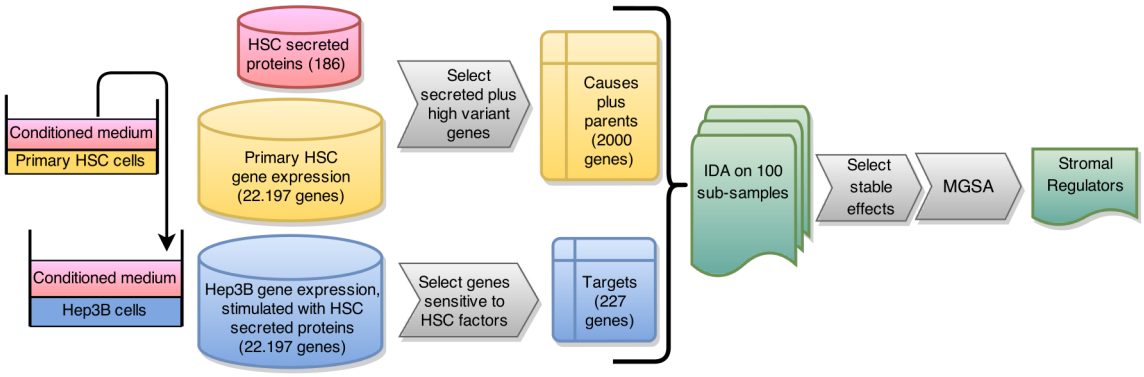


Figure 1: **Overview of the experimental and computational approach to identify secreted stromal (HSC) factors which influence tumor (HCC) gene expression.** Conditioned medium of primary human HSC (n=15) was transferred onto human Hep3B HCC cells. Gene expression data of HSC and HCC cells were filtered to reduce the dimensionality of the data and to build cause-and-effect (target) matrices. The matrices served as input for the IDA algorithm which estimates causal effects for each cause on each target gene. Causal effects that were stable across sub-sampling runs (i.e., that were stable with respect to small perturbations of the data) were retained and subjected to Model-based Gene Set Analysis (MGSA) to extract a sparse set of HSC genes influencing HCC cell gene expression.

We modified this algorithm in such a way that it covered the list of cancer genes responsive to stromal factors with the sets of HSC targets predicted by IDA. Thus HSC genes were in competition to each other: an analysis based on frequencies (how many HCC genes does each HSC gene affect) discovers redundant HSC genes that target the same HCC genes. Our approach strove for a maximum coverage of the target genes with a minimum number of HSC secreted genes. We identified 10 HSC secreted proteins which covered the majority of gene expression changes observed in HCC cells. The list consisted of PGF, CXCL1, PAPPA, IGF2, IGFBP2, POSTN, NPC2, CTSB, HGF, and CSF1. Notably, the set of the most influential HSC regulators included several well-known tumor-promoting genes such as placental growth factor (PGF) [PWB$^+$05], and the chemokine CXCL1, which promotes HCC angiogenesis and growth [TML$^+$12]. Periostin (POSTN) is a secreted cell adhesion protein whose expression levels are directly related to metastatic potential and poor prognosis of HCC [LWJ$^+$13]. High expression levels of the macrophage colony-stimulating factor 1 (CSF1) are another indicator of tumor progression and poor survival in HCC patients [BFY$^+$06]. Over-expression of cathepsin B (CTSB), on the other hand, promotes HCC cell migration and invasion [CCJ$^+$12].

PAPPA is a novel stromal factor which activates NFkB signaling in cancer cells

In our paper [EAOR$^+$15], we identified PAPPA as a novel stromal regulator of HCC cell gene expression. As many of the cancer genes that changed gene expression levels upon incubation with stroma-conditioned medium are NFkB pathway members or targets of the transcription factor NFkB, we experimentally tested whether PAPPA could induce NFkB activity. Indeed, recombinant PAPPA protein together with conditioned medium induced a stronger induction of NFkB signaling than conditioned medium alone. We could also show that PAPPA is solely secreted by HSCs and not by HCC cells, and its protein levels correlate with fibroblast markers in patient samples, indicating that stromal cells are the major source of PAPPA also in HCC tissue. Finally, we could show that increased levels of PAPPA indicate advanced stage HCC in clinical samples.

PeerJ PrePrints

3 Presentation outline

The presentation will first motivate and introduce the importance of cell communication of different cell types in tumor tissue. Then the experimental setting to produce systems-wide unidirectional cell communication data will be presented. The main part of the presentation will focus on the computational model to derive the most important stromal factors which influence tumor cell gene expression. The last part will briefly highlight biological findings with this approach and their clinical implications.

References

[BFY+06] Anuradha Budhu, Marshonna Forgues, Qing-Hai Ye, Hu-Liang Jia, Ping He, Krista A. Zanetti, Udai S. Kammula, Yidong Chen, Lun-Xiu Qin, Zhao-You Tang, and Xin Wei Wang. Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. *Cancer Cell*, 10(2):99–111, Aug 2006.

[BRG11] Sebastian Bauer, Peter N Robinson, and Julien Gagneur. Model-based gene set analysis for Bioconductor. *Bioinformatics*, 27(13):1882–1883, Jul 2011.

[CCJ+12] Wan-Nan Chen, Jin-Yan Chen, Bo-Yan Jiao, Wan-Song Lin, Yun-Li Wu, Ling-Ling Liu, and Xu Lin. Interaction of the hepatitis B spliced protein with cathepsin B promotes hepatoma cell migration and invasion. *J Virol*, 86(24):13533–13541, Dec 2012.

[EAOR+15] Julia C. Engelmann, Thomas Amann, Birgitta Ott-Rötzer, Margit Nützel, Yvonne Reinders, Jörg Reinders, Wolfgang E. Thasler, Theresa Kristl, Andreas Teufel, Christian G. Huber, Peter J. Oefner, Rainer Spang, and Claus Hellerbrand. Causal Modeling of Cancer-Stromal Communication Identifies PAPPA as a Novel Stroma-Secreted Factor Activating NFkB Signaling in Hepatocellular Carcinoma. *PLoS Comput Biol*, 11(5):e1004293, May 2015.

[JQF+09] Min-Jie Ju, Shuang-Jian Qiu, Jia Fan, Yong-Sheng Xiao, Qiang Gao, Jian Zhou, Yi-Wei Li, and Zhao-You Tang. Peritumoral activated hepatic stellate cells predict poor clinical outcome in hepatocellular carcinoma after curative resection. *Am J Clin Pathol*, 131(4):498–510, Apr 2009.

[LWJ+13] Yang Lv, Wei Wang, Wei-Dong Jia, Qi-Kai Sun, Jian-Sheng Li, Jin-Liang Ma, Wen-Bin Liu, Hang-Cheng Zhou, Yong-Sheng Ge, Ji-Hai Yu, Hong-Hai Xia, and Ge-Liang Xu. High-level expression of periostin is closely related to metastatic potential and poor prognosis of hepatocellular carcinoma. *Med Oncol*, 30(1):385, Mar 2013.

[MKB09] Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

[Pea00] Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge Univ. Press, Cambridge, Cambridge, 2000.

[Pea03] Judea Pearl. Statistics and causal inference: A review. *Test*, 12:281–318, 2003.

[PWB+05] Christian Parr, Gareth Watkins, Mike Boulton, Jun Cai, and Wen G. Jiang. Placenta growth factor is over-expressed and has prognostic value in human breast cancer. *Eur J Cancer*, 41(18):2819–2827, Dec 2005.

[SGS00] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, Cambridge, 2nd edition, 2000. ISBN 0-262-19440-6.

[TC06] Thea D. Tlsty and Lisa M. Coussens. Tumor stroma and regulation of cancer development. *Annu Rev Pathol*, 1:119–150, 2006.

[TML+12] Kwan Ho Tang, Stephanie Ma, Terence K. Lee, Yuen Piu Chan, Pak Shing Kwan, Carol M. Tong, Irene O. Ng, Kwan Man, Ka-Fai To, Paul B. Lai, Chung-Mau Lo, Xin-Yuan Guan, and Kwok Wah Chan. CD133+ liver tumor-initiating cells promote tumor angiogenesis, growth, and self-renewal through neurotensin/interleukin-8/CXCL1 signaling. *Hepatology*, 55(3):807–820, Mar 2012.

[WF14] Michael C. Wallace and Scott L. Friedman. Hepatic fibrosis and the microenvironment: fertile soil for hepatocellular carcinoma development. *Gene Expr*, 16(2):77–84, 2014.

[ZZY+11] Wenxiu Zhao, Lei Zhang, Zhenyu Yin, Weixue Su, Guangli Ren, Changsheng Zhou, Junyong You, Jia Fan, and Xiaomin Wang. Activated hepatic stellate cells promote hepatocellular carcinoma development in immunocompetent mice. *Int J Cancer*, 129(11):2651–2661, Dec 2011.

PeerJ PrePrints

Integrative Analysis of Epigenomics Data using the bidirectional hidden Markov Model in the R package STAN

Benedikt Zacher[1,2], Rafael Campos-Martin[1,3], Julien Gagneur[2], Achim Tresch[1,2,3,*]
[1]*Department of Biology, University of Cologne;* [2]*Gene Center, Ludwig-Maximilians-University Munich;* [3]*Max-Planck Institute for Plant Breeding Research, Cologne*
[*]send correspondence to: tresch@mpipz.mpg.de

PeerJ PrePrints

Introduction

Current sequencing-based experimental techniques like RNA-seq and ChIP-Seq generate a wealth of data which can be aligned to the genome of the targeted organism. Yet the integrated analysis of multiple such datasets requires methods for their automated and efficient statistical analysis. One major goal is to annotate the genome, i.e., to cluster genomic positions into functional groups based on the observations made at these positions. One might, e.g., want to dissect the process of RNA trancription into distinct phases characterized by the presence of different protein complexes that change their composition as the RNA Polymerase moves along the DNA. Ideally, such a clustering accounts for the dependency of observations induced by the linear structure of the DNA and the processes associated to it. Hidden Markov models (HMMs) have been used extensively to partition the genome into discrete functional states that can be interpreted as DNA-associated protein complexes. They have been used to infer chromatin states, and annotate enhancers, promoters and transcribed and quiescent regions in human [TDNS07, EK12] and fly [FvBB$^+$10].

Current HMM-based approaches ignore the fact that DNA-related processes may occur in forward or reverse direction. [KGP14] use time-reversible Markov chains to alleviate this drawback. Still, this model is not able to infer the diretionality of DNA-related processes, nor do they properly integrate strand specific (e.g., RNA expression) with non-strand-specific (e.g., ChIP) data. In order to address these points, our present contribution highlights the bidirectional hidden Markov model (bdHMM) introduced in [ZLC$^+$14].

Results

The main idea of the bdHMM is to have so-called twin states, one for each strand and genomic state. Transitions between twin states are coupled by a generalized time-reversibility condition, which replaces the ordinary time-reversibility constraint for reversible HMMs (see Methods for a precise definition). bdHMMs can identify forward and reverse directed states by taking into account directional information contained in each single observation. We derived an efficient analog of the Baum-Welch expectation-maximization (EM) algorithm for bdHMM parameter learning. The bdHMM model along with the EM algorithm is implemented in the open source R/Bioconductor package STAN [ZGT14]. STAN allows the modeling of multivariate Gaussian, Poisson, negative binomial, and multinomial emission distributions and arbitrary independent combinations thereof. It thereby provides a general and flexible framework for obtaining a directed functional state annotation from genomics data.

We applied the bdHMM to a combined RNA transcription and ChIP data set of RNA Polymerase II-associated general transcription factors in yeast. The bdHMM annotated the genome with transcription states (Figure 1), which were characterized by different compositions of the Polymerase II complex. Searching this sequence of states with regular expressions recovers the majority of transcribed loci. We reveal gene-specific variations in the yeast transcription cycle and we find an alternative transcription termination pathway for antisense transcripts. Application of the bdHMM to chromatin modification

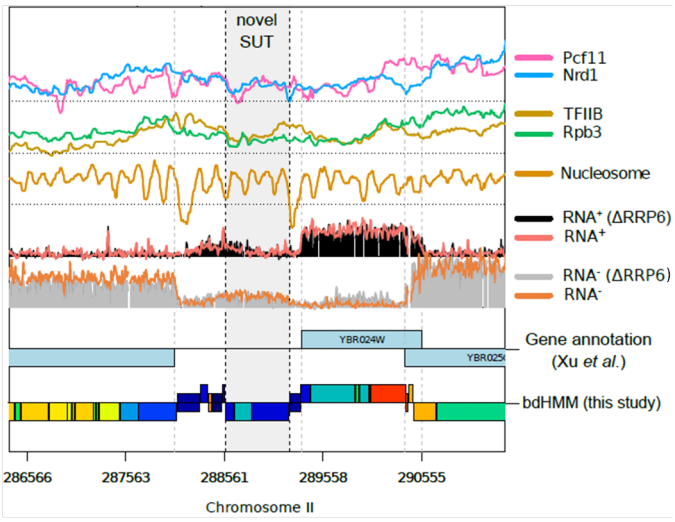


Figure 1: De novo transcript annotation by the bdHMM. Top panels: The data which was used to train the bdHMM include termination factors Pcf11 (pink) and Nrd1 (blue), initiation factor TFIIB (ocre), the RNA Polymerase II subunit Rpb3 (green), Nucleosomes (orange), and strand-specific RNA-Seq expression data in wild type cells (dark and light red) and cells deficient for the nuclear exosome (black and grey). Middle panel: The transcript annotation by [XWG+09], which was not known to the bdHMM, was used as a gold standard. Bottom panel: Viterbi path derived from the bdHMM. Different colors indicate different states. States above (below) the baseline indicate reverse (forward) states, the other states are undirected. The grey area highlights a novel SUT (Stable unannotated transcript, a stable non-coding RNA) region predicted to be expressed on the + strand by the bdHMM yet not captured by former annotations based on the wild-type RNA levels alone. (Modified after [ZLC+14])

data in human T cells provides evidence for existence of directed chromatin state patterns around transcribed regions in the human genome.

Methods

A hidden Markov model is a tuple $\theta = (\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ such that $\mathcal{K}$ is a finite state set, $\pi = (\pi_i)_{i\in\mathcal{K}}$ is the initial state distribution, $A = (a_{ij})_{i,j\in\mathcal{K}}$ is a $\mathcal{K} \times \mathcal{K}$ transition matrix, and $\Psi = \{\psi_i;\ i \in \mathcal{K}\}$ is a set of probability distributions on the observation space $\mathcal{D}$. An HMM defines a probability distribution on a sequence of observations $\mathcal{O} = (o_1, ..., o_T)$. Each observation o_t is *emitted* by a corresponding hidden (unobserved) state variable s_t which can assume values in $\mathcal{K}$. The value of s_t determines the probability of observing o_t by $\Pr(o_t \mid s_t) = \psi_{s_t}(o_t)$. The hidden variables are assumed to form a homogenous Markov chain $\mathcal{S} = (s_1, ..., s_T)$ with time-independent transition probabilities $\Pr(s_t = j \mid s_{t-1} = i) = a_{ij}$, $i, j \in \mathcal{K}$, $t = 2, ..., T$, and with initial state distribution $\Pr(s_1 = i) = \pi_i$, $i \in \mathcal{K}$. The full likelihood of an HMM is

$$\begin{aligned}\Pr(\mathcal{O}, \mathcal{S};\, \theta) &= \Pr(\mathcal{S};\, \theta) \cdot \Pr(\mathcal{O} \mid \mathcal{S};\, \theta) = Pr(s_1;\, \pi) \cdot \prod_{t=2}^{T} Pr(s_t \mid s_{t-1};\, A) \cdot \prod_{t=1}^{T} Pr(o_t \mid s_t;\, \Psi) \\ &= \pi_{s_1} \cdot \prod_{t=2}^{T} a_{s_{t-1}s_t} \cdot \prod_{t=1}^{T} \psi_{s_t}(o_t)\end{aligned}$$

Given a sequence of observations $\mathcal{O}$, the Viterbi algorithm can be used to find the maximum likelihood hidden state sequence $\mathcal{S}$, thus assigning to each position a state in $\mathcal{K}$. This Viterbi path is commonly

used as annotation of the genome (Figure 2a,b). The main idea of the bdHMM is to split the state space $\mathcal{K}$ into undirected states, and pairs of directed (forward and reverse) twin states. Symmetry conditions couple the emission and transition probabilities of twin states in a meaningful way (Figure 2a,c).

Definition. A **bidirectional hidden Markov model** (bdHMM) is a tuple $\theta = ((\mathcal{K}, \kappa), \pi, A, (\mathcal{D}, \delta), \Psi)$ such that $(\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ is an HMM. Additionally, $\kappa : \mathcal{K} \to \mathcal{K}, k \mapsto \bar{k}$ and $\delta : \mathcal{D} \to \mathcal{D}$, $o \mapsto \bar{o}$ are involutions ($\kappa^2 = \mathrm{id}$, $\delta^2 = \mathrm{id}$). The involustion κ defines the directed twin states by mapping a state j to its direction-reversed twin state $\bar{j}$, while leaving undirected states fixed. The involution δ maps an observation o to $\bar{o}$ by swapping strand-specific observations. Finally, the following symmetry conditions hold:

1. Generalized detailed balance relation: The transition matrix A and the initial state distribution π satisfy
$$\pi_i a_{ij} = \pi_{\bar{j}} a_{\bar{j}\bar{i}} \quad , \; i, j \in \mathcal{K} \tag{1}$$
2. Initiation symmetry: The initial state distributionπ satisfies
$$\pi_i = \pi_{\bar{i}} \quad , \; i \in \mathcal{K} \tag{2}$$
3. Observation symmetry: Ψ satisfies
$$\psi_i(o) = \psi_{\bar{i}}(\bar{o}) \quad , \; i \in \mathcal{K}, \, o \in \mathcal{D} \tag{3}$$

Why did we specifically choose conditions (1)-(3) as the defining properties of a bdHMM? To motivate our definition, we give an alternative characterization of the bdHMM in terms of a biologically motivated condition. It is natural to require that a directionality-aware HMM marginally cannot distinguish between a forward transition i, j from position $t-1$ to t when observing x, y at the corresponding positions, and the reverse transition $\bar{j}$, $\bar{i}$ at position $t-1$ to t when observing $\bar{y}$, $\bar{x}$ at the corresponding positions (Figure 2d). In other words, we require that

$$\Pr(s_{t-1} = i, s_t = j, o_{t-1} = x, o_t = y; \, \theta) = \Pr(s_{t-1} = \bar{j}, s_t = \bar{i}, o_{t-1} = \bar{y}, o_t = \bar{x}; \, \theta) \tag{4}$$

holds for all $i, j \in \mathcal{K}$, $x, y \in \mathcal{D}$, $t = 1, 2, \ldots$. Under very mild additional assumptions that are always met in practice, this condition characterizes a bdHMM:

Theorem. Let $\theta = ((\mathcal{K}, \kappa), \pi, A, (\mathcal{D}, \delta), \Psi)$ be a tuple with involutions $\kappa : \mathcal{K} \to \mathcal{K}$, $\delta : \mathcal{D} \to \mathcal{D}$, such that $(\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ is an HMM. Let each $\psi_i \in \Psi$ be non-constant, and and let A be irreducible, i.e., there exists some positive integer r such that $(A^r)_{ij} > 0$ for all $i, j \in \mathcal{K}$. Then, θ is a bdHMM if and only if Condition (4) holds.

Discussion

Bidirectional Hidden Markov Models (bdHMMs), are a novel method for de novo and unbiased inference of directed genomic states from genome-wide profiling data. It allows for the integration of strand-specific data such as RNA expression together with non-strand-specific data such as ChIP occupancy. It can jointly model nominal, continuous and count data by a variety of emission distributions. The open-source package STAN provides a fast, multiprocessing implementation that can process the human chromatin data set in less than one day using a 20 CPU compute cluster. The most significant advance of bdHMM analysis over previous methods is its potential to de novo identify characteristic sequences (patterns) of directed states on the genome. The explicit modeling of forward and reverse states detected an alternative transcription termination pathway which is primarily associated with antisense transcripts. We find that directed patterns of histone modifications are ordered according to the direction of RNA transcription. A bdHMM has essentially the same number of parameters as a comparable standard HMM, and its learning is done at the same speed. Thus, the inference of directionality is without additional costs. We therefore expect the bdHMM to have a broad range of applications in genomics and epigenomics.

PeerJ PrePrints

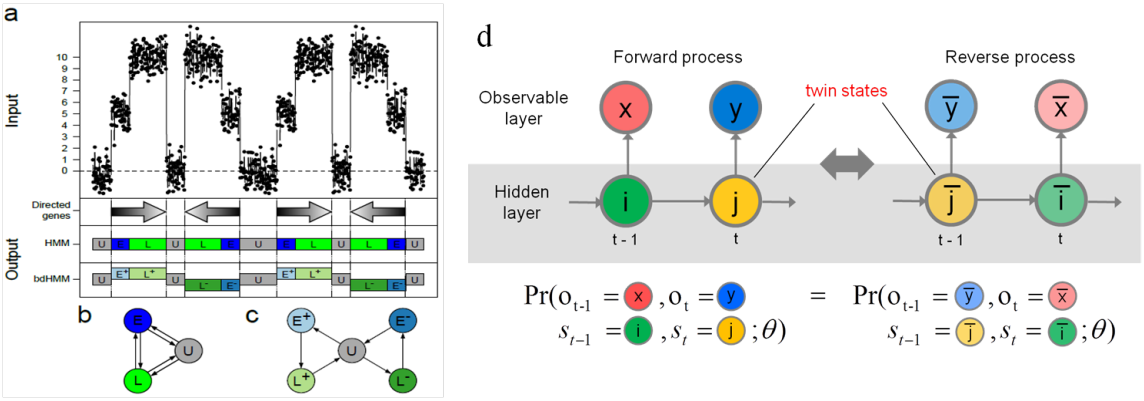


Figure 2: Toy example of a bdHMM model. (a) Simulated occupancy signal (1st track from the top) for a putative factor with a low level (centered at 0) in untranscribed regions (state U), an intermediate level in 5' part of genes (state E), and a high level in 3' part of genes (state L). Arrows (2nd track) depict boundaries and orientation of transcription. Unlike standard HMMs (3rd track) bdHMM (4th track) infer strands (+ or -) to expressed states (E, L). (b) HMM transition graph. Because orientation of transcription is not modeled by standard HMMs, the spurious reverse transitions (E⇒ U, L⇒ E, and U⇒ L) are as likely as the correctly oriented transitions (U⇒ E, E⇒ L, and L⇒ U). (c) bdHMM transition graph. In contrast to HMMs, bdHMMs explicitly model strand-specific expression states (E^+/E^- and L^+/L^-), which results in the correct inference of oriented transitions. (d) Illustration of condition (4), the defining property of a bdHMM. (Modified after [Zacher et al. 2014])

References

[EK12] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9(3):215–216, Mar 2012.

[FvBB+10] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, and B. van Steensel. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, 143(2):212–224, Oct 2010.

[KGP14] David Knowles, Zoubin Ghahramani, and Konstantina Palla. A reversible infinite HMM using normalised random measures. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1998–2006, 2014.

[TDNS07] R. E. Thurman, N. Day, W. S. Noble, and J. A. Stamatoyannopoulos. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.*, 17(6):917–927, Jun 2007.

[XWG+09] Z. Xu, W. Wei, J. Gagneur, F. Perocchi, S. Clauder-Munster, J. Camblong, E. Guffanti, F. Stutz, W. Huber, and L. M. Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037, Feb 2009.

[ZGT14] Benedikt Zacher, Julien Gagneur, and Achim Tresch. STAN: STrand-specific ANnotation of genomic data. R package version 1.2.0. *Bioconductor 3.0*, 2014.

[ZLC+14] Benedikt Zacher, Michael Lidschreiber, Patrick Cramer, Julien Gagneur, and Achim Tresch. Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. *Molecular systems biology*, 10(12):768, 2014.

PeerJ PrePrints